Beyond Formants: Vowel Perception at High Fundamental Frequencies

Thesis (cumulative thesis)
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by Daniel Friedrichs

Accepted in the fall semester 2016
on the recommendation of the doctoral committee:
Prof. Dr. Volker Dellwo (main supervisor)
Prof. Dr. Martin Meyer

Zurich, 2017

Abstract

It is commonly assumed that the vowel identification process is mainly driven by an underlying acoustic representation of formant frequency patterns. This assumption contributes largely to the pervasive idea that listeners' ability to recognize vowels has to be poor at very high fundamental frequencies (f_o) due to a sparse sampling of the vocal tract transfer function. In this cumulative thesis, it is shown that the phonological function of vowels can be maintained at f_os up to at least 880 Hz and that listeners can identify the point vowels /i a u/ at even higher f_os. Auditory excitation patterns revealed highly differentiable representations for these categories that can be used as landmarks for vowel category perception at high f_os. The results suggest that theories of vowel perception based on overall spectral shape will provide a fuller account of vowel perception than those based solely on formant frequency patterns.

Zusammenfassung

Es ist eine gemeinhin akzeptierte Annahme, dass die Vokalperzeption von der akustischen Repräsentation der Formantfrequenzmuster abhängig ist. Diese Auffassung trägt entscheidend zu der weit verbreiteten Einschschätzung bei, dass menschliche Vokalproduktionen auf hohen Grundfrequenzen (f_o) ihre Verständlichkeit verlieren müssen, da die Transferfunktion des Vokaltraktes nicht ausreichend abgetastet werden kann und somit die auf niedrigeren f_o üblichen Muster nicht mehr im Frequenzspektrum vorzufinden sind. In dieser kumulativen Dissertation wird gezeigt, dass die phonologische Funktion von Vokalen bis zu einer f_o von mindestens 880 Hz erhalten bleiben kann und dass Hörer die Vokale /i a u/ selbst auf darüber hinausgehenden f_o identifizieren können. Anhand auditorischer Erregungsmuster konnten zudem leicht differenzierbare Repräsentationen dieser Kategorien nachgewiesen werden, welche auf hohen f_o als Orientierungshilfe bei der Perzeption von Vokalkategorien genutzt werden dürften. Die Ergebnisse lassen den Schluss zu, dass eine Theorie der Vokalperzeption, welche auf einer umfänglichen spektralen Form basiert, einen geeigneteren Ansatz darstellt als eine solche, die lediglich auf Formantfrequenzmustern beruht.

Acknowledgements

First and foremost, I would like to thank my main advisor Volker Dellwo. He helped me greatly to take my first steps into the fascinating world of science. I am also very thankful to my other advisors and collaborators Dieter Maurer, Martin Meyer, and Stuart Rosen. They have supported me in many ways, and I truly enjoyed working with them.

My time at the University of Zurich was made very enjoyable due to many dear colleagues and friends. I thank my brewing companion Steve Moran for exploring with me the secrets of another IPA, and Sandra Schwab for being so patient with me during our French lunch breaks (*merci mille fois!*). I would also like to thank my lab mates Lei He, Kostis Dimos, and Thayabaran Kathiresan. It was always fun and I truly enjoyed working with you guys. I also owe a lot to my friends outside the lab who greatly helped to make life besides *formants* memorable too. Thank you, Hendrik, André, Louise, Pierre, Johan, and all the others for the good times we had together in the last years.

My family in Germany deserves my deep gratitude for always supporting and believing in me. Sincere thanks also go to all the members of the Maeder-Ingvar family who opened their home to me and made me feel welcome in Switzerland.

Last, but by no means least, I am incredibly indebted to Cecilia Maeder for her unbelievable support and encouragement. It will be your turn soon!

Table of Contents

List of Tables

List of Figures

Close your eyes, prick up your ears, and from the softest sound to the wildest noise, from the simplest tone to the highest harmony, from the most violent, passionate scream to the gentlest words of sweet reason, it is but *Nature* who speaks, revealing her being, her power, her life, and her relatedness, so that a blind person, to whom the infinitely visible world is denied, can grasp an infinite vitality in what can be heard.

Johann Wolfgang von Goethe

CHAPTER 1
Synopsis

1.1 Conceptual overview

The main body of this cumulative thesis consists of four peer-reviewed papers on vowel perception at high fundamental frequencies (study I, II, III) and methodological problems that go along with the acoustic analysis of steady-state vowels (study IV). In this synopsis, the underlying theoretical framework of the dissertation is introduced, and the four studies are briefly summarized. In addition, an outlook on future investigations is provided that are required to fully understand the results obtained in the studies I, II, and III.

For clarity's sake, a consistent layout and consecutive numbering of the sections, figures and tables are used throughout this work. Additional changes have been applied to the symbolic notation of the terms *fundamental frequency*, *formant frequency*, *resonant frequency*, and *harmonic* following the suggestions of a group of voice scientists recently published in the *Journal of the Acoustical Society of America* (Titze et al., 2015). More detailed information on key terms and abbreviations can be found in the Appendix.

1.2 Research background

Listeners' ability to identify vowels is commonly understood against the background of the *source-filter theory of speech production* (Chiba and Kajiyama, 1941; Stevens and House, 1955; Fant, 1960; see also Eq. 1.1, and Fig. 1–1). According to this theory, the *source* signal of a voiced speech sound, and thus a vowel, is produced when the vocal folds inside the larynx are set into vibration by an air stream that is expelled from the lungs. This creates a complex acoustic waveform that is called a glottal pulse and whose shape is similar to that of a sawtooth wave showing a moderate increase in pressure amplitude followed by a sharp drop. The number of glottal pulses per second (i.e., the periodicity) determines the fundamental frequency (f_o) in human speech. A Fourier transform of this source signal reveals a power spectrum with harmonics typically decaying by about 10 to 15 decibel in sound pressure level (dB SPL)—depending on the sub-glottal pressure that is applied to the vocal folds (Titze, 2015)—for every octave the frequency increases.

The buzzing-like sound then travels through the vocal tract, which is considered to be a linear time-invariant system (Stevens, 2000). The frequency response of this *filter* is called the vocal tract's transfer function. It is determined by its length and shape, which both can be altered by the talker, for example, by raising the larynx and changing the position of the tongue. When the modified acoustic signal finally radiates from the mouth, the efficiency of the sound transmission increases at a rate of about 6 dB SPL per octave at frequencies within the range of about 300–4000 Hz (Diehl, 2008; Stevens, 2000). As a product, the source function $S(f)$, the vocal

tract's transfer function T(f), and the radiation characteristic from the mouth R(f) have a specific output spectrum, which is given by

$$p_r(f) = S(f)T(f)R(f), \tag{1.1}$$

where r is the distance from the mouth.

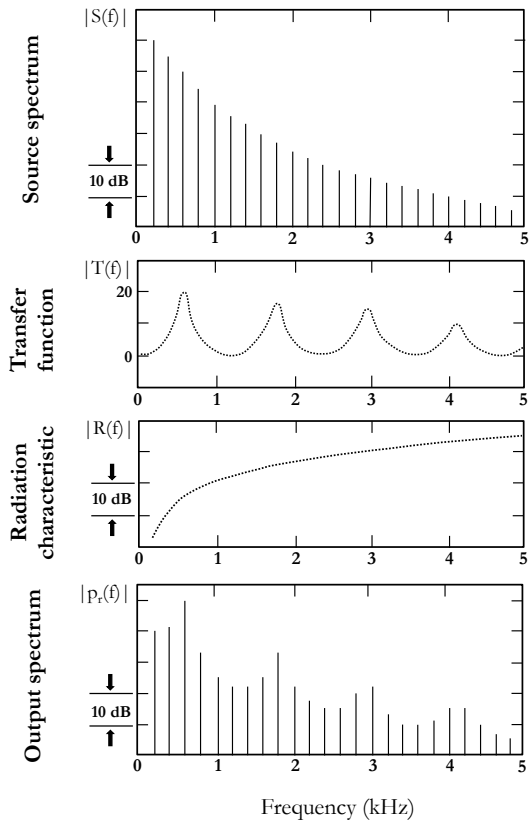

Figure 1–1: Schematic illustration of the *source-filter theory of speech production* with the individual components. The output spectrum (bottom) is the product of the source spectrum, the vocal tract transfer function, and the radiation characteristic from the mouth. (Adapted from Stevens, 2000)

The output spectrum exhibits distinctive patterns of absolute or relative spectral maxima. These are known as *formants* and the frequency at the maxima is called *formant frequency* (ANSI/ASA S1.1-2013, p. 62). Among phoneticians, formant frequency patterns—in particular, the distribution of the first three formant frequencies (F_1–F_3)—are commonly assumed to be the most salient acoustic cues to vowel perception (e.g., Rakerd and Verbrugge, 1984; Terbeek and Harshman, 1972; Shepard, 1972; Pols et al., 1969).

Therefore, it is not surprising that most influential works on the acoustic characteristics of vowels used formant frequency measurements to describe the differences between vowel categories in individual languages (e.g., Assmann and Katz, 2000; Pätzold and Simpson, 1997; Yang, 1996; Hillenbrand et al., 1995; Hagiwara, 1997; Peterson and Barney, 1952). This seems reasonable as it has also been shown numerous times that vowels (Bladon and Fant, 1978) and voiced speech in general (Remez et al., 1981) can be synthesized in an intelligible form on the basis of only the first two or three formant frequencies. Table 1–1 provides an example of such a data collection, which can be found in an often-cited paper by Hillenbrand et al. (1995) on the acoustic characteristics of American English vowels.

		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɜ/
f_o	M	138	135	129	127	123	123	121	129	133	143	133	130
	W	227	224	219	214	215	215	210	217	230	235	218	217
	C	246	241	237	230	228	229	225	236	243	249	236	237
F_1	M	342	427	476	580	588	768	652	497	469	378	623	474
	W	437	483	536	731	669	936	781	555	519	459	753	523
	C	452	511	564	749	717	1002	803	597	568	494	749	586
F_2	M	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
	W	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588
	C	3081	2552	2656	2267	2501	1688	1210	1137	1490	1345	1546	1719
F_3	M	3000	2684	2691	2605	2601	2522	2538	2459	2434	2343	2550	1710
	W	3372	3053	3047	2979	2972	2815	2824	2828	2827	2735	2933	1929
	C	3702	3403	3323	3310	3289	2950	2982	2987	3072	2988	3145	2143

Table 1–1: Average fundamental frequencies and formant frequencies of 12 vowels of American English produced by 45 men, 48 women, and 46 children (27 boys, 19 girls; ten- to 12-year old). The vowels were produced by 139 talkers in /hVd/ context resulting in the words "heed", "hid", "hayed", "head", "had", "hod", "hawed", "hoed", "hood", "who'd", "hud", "heard", "hoyed", "hide", "hewed", and "how'd". 20 phonetically trained listeners identified all vowels with an error rate not greater than 15%. All measurements are in Hz and were made in a 56 ms steady-state portion of the respective vowel. (Adapted from Hillenbrand et al., 1995)

However, some problems that are associated with describing vowels solely on the basis of their formant frequency patterns have been discussed over the last decades.

For example, an ambiguity of F_1–F_2 combinations for different vowels between talkers was found in most of the above-mentioned studies (see Fig. 1–2; for an overview on this matter, see the introduction in Maurer et al., 2000). The fact that different vowel categories between talkers can have very similar formant frequency patterns which do not necessarily lead to a loss of vowel category information is often explained by either the role of vowel inherent spectral change (VISC) (Nearey and Assmann, 1986) or with the help of talker normalization processes (for an overview, see Adank et al., 2004). VISC addresses the fact that formant frequency trajectories are normally not entirely flat due to the highly dynamic process of articulation. It has been shown in several studies (see Morrison and Assmann, 2012, for a collection of papers on this topic) that spectral variability within a vowel has a substantial effect on listeners' perception. Thus, it seems plausible that ambiguous formant frequency patterns might only be a result of simplifying the acoustic features of vowels by using average values instead of ranges. Normalization theories, on the other hand, are driven by the assumption that listeners need to adapt their perception continuously as they have to deal with a high variability of acoustic inputs representing only a relatively small number of phonological categories when they listen to different or even to single talkers. Joos (1948), for example, argued that listeners could create idiosyncratic vowel spaces for individual talkers based on the first utterances they hear from them. More recent approaches are built on a theory of template-matching processes, in which the templates are either based on distinctive acoustic features (i.e., a *feature- or contrast-based model*; see Liberman and Mattingly, 1989; Stevens, 2002) or entire tokens that listeners have encountered during their life (i.e.,

an *exemplar-based model*; see Johnson, 1997; Pierrehumbert, 2001). In this context, secondary cues (e.g., duration, formant transitions, intensity levels) within the vowel sound or in its environment are also often considered to help listeners to distinguish between categories.

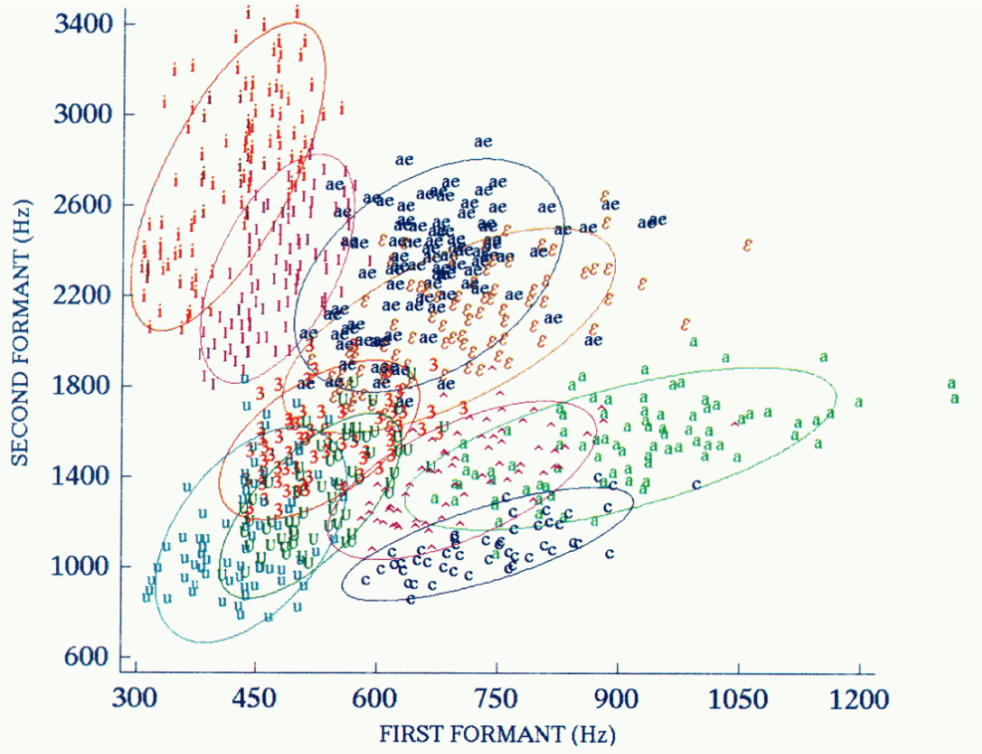

Figure 1–2: Scatter plot showing formant frequency measurements of 12 vowels of American English in an $F1$-$F2$ space. Different vowel categories overlap in this two-dimensional space, mainly due to between-talker differences. (Reproduced from Hillenbrand et al., 1995)

Although it seems plausible that normalization procedures and VISC might provide an answer to the questions arising from ambiguous formant frequency patterns,

they do not consider the possibility that other invariant spectral cues than formants could also play an important role in human vowel perception. Support for this assumption can be found in everyday-life speech communication when f_o is relatively high. Given that patterns of formant frequencies are the most salient cues to vowel perception, listeners' ability to recognize vowels should be poor at very high f_os due to a sparse sampling of the vocal tract transfer function. This holds true, in particular, when the normal range of the first formant frequency (F_1) is exceeded by f_o, and the higher formants are poorly specified due to a wide spacing of the harmonics (see Fig. 1–3). Although it is obvious that this is not the case and recognizable utterances at very high pitches can be found (e.g., in infant-directed speech, Trainor and Desjardins, 2002), all influential studies on the acoustic characteristics of vowels are based on samples with relatively low f_os. This might be the case as (a) it guarantees a sufficient sampling of the vocal tract's transfer function, which should lead to better measurements, (b) laboratory speech does not show a high degree of f_o variability, (c) the standard analysis tools for formant measurements (e.g., *linear predictive coding (LPC)* which will be explained in more detail in chapter 5, cannot be used for acoustic signals with very high f_os. (e.g., Monsen and Engebretson, 1983 found that the accuracy of LPC decreases largely at f_os above 350 Hz), and (d) it seems to be commonly assumed by phoneticians and acousticians that vowel identification has to decrease with an increasing f_o.

The latter assumption is supported by numerous findings in the field of singing research, and solely open vowels like /a/ and /ɑ/, which typically show the highest F_1 of all vowels, were found to be identifiable at the highest notes (i.e., in the f_o range

around 1 kHz) (Deme, 2014; Gregg and Scherer, 2006; Benolken and Swanson, 1990; Scotto di Carlo and Germain, 1985; Sundberg and Gauffin, 1982; Mozorov, 1965; Howie and Delattre, 1962).

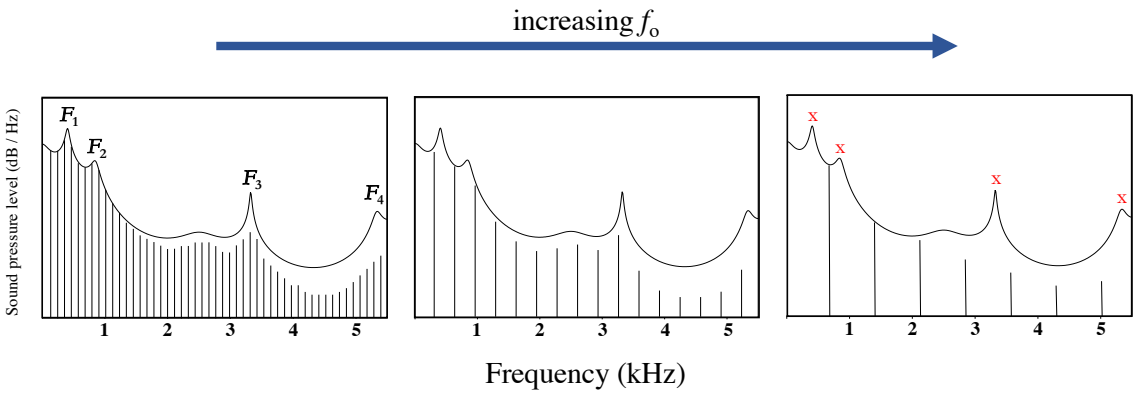

Figure 1–3: Schematic illustration of an undersampling of the vocal tract transfer function as a result of an increasing fundamental frequency. The figure shows a typical transfer function of the vowel /u/ of an adult male talker superimposed on spectra with increasing spacing of the harmonics. This leads increasingly to a loss of formant pattern information and should, thus, inevitably compromise vowel recognition at very high f_os.

It seems likely, however, that the results observed in singing research were strongly influenced by articulatory and acoustical adjustments applied by singers (in particular, by operatic singers, which build the main body of subjects in the studies mentioned above). In experimental studies such as Joliveau et al. (2004) it has been shown, for example, that sopranos shift the first resonant frequency (f_{R1}) of their vocal tract – and thus F_1 – to the vicinity of f_o as soon as f_o drastically exceeds the normal range of f_{R1} of an intended vowel. This tuning of f_{R1} is achieved

by increasing the jaw opening and reducing the maximum constriction of the vocal tract (Sundberg, 1975; Sundberg, 2013). As f_o gains considerable amplitude when being close to a resonant frequency, these maneuvers may help a singer to maintain vocal power and timbral homogeneity (Smith and Wolfe, 2009). However, the acoustic modifications associated with shifting a resonant frequency may lead to different spectral patterns and consequently to a confusion of vowel categories.

Given this situation, it is surprising that few studies have investigated vowel recognition outside Western operatic singing at very high f_os as there is evidence that even a sparsely sampled vocal tract transfer function still carries information, which can be used by listeners to recognize different vowels, despite a likely absence of the supposed F_1 and an undersampling of the higher formants. Smith and Scott (1980), for example, reported listeners' identification performance significantly above chance level (mean of 70% correct) for the four front vowels /i ɪ ɛ æ/, which were produced by a soprano in isolation at an f_o of about 880 Hz (i.e., the musical note A5) with a raised larynx (i.e., a shortened vocal tract), and thus not in an articulation mode typical for Western operatic singers. When asked to produce the same vowels in her operatic singing style, identification dropped to a mean of 4% correct at the same f_o. Maurer and Landis (1996) showed that infant and adult talkers can produce identifiable versions of the vowels /i a o u/ but not of /e/ at an f_o between about 500–870 Hz that was individually chosen by the talker. The reason why little attention has been paid to these studies so far might be that both studies did not consider several factors that could have had an influence on listeners' identification performance, for example, vowel intrinsic duration (Hillenbrand et al., 2000), vowel

intrinsic intensity (Lehiste and Peterson, 1959), and frication noise shaped by the co-articulated vocal tract resonances of an intended vowel (Schnupp et al., 2011:37).

The present dissertation, however, follows up on these findings. The studies I, II, and III address the question whether all long vowels of a language (German) can remain identifiable when they are produced and perceived with quasi-flat f_o contours and resonance trajectories at very high f_os and in different experimental conditions. In addition, stimuli sets have been chosen that allow controlling for the influence of possible secondary cues. Study IV addresses methodological issues that go along with the acoustic analysis of steady-state vowels, in particular, those that are produced at high f_os.

1.3 Study I

Study I provides evidence that the phonological function of vowels can be maintained at fundamental frequencies up to 880 Hz. A female talker with professional voice training (a German native talker and trained Musical-Theatre singer) produced the eight vowels /i y e ø ɛ a o u/ within minimal pairs with the stimulus vowel in contrastive position at nine f_os between 220–880 Hz. The talker was instructed to focus exclusively on producing recognizable vowels in a speech-like mode and to ignore typical voice aesthetics of her singing style. In a binary forced-choice task, two groups of native German listeners (each N=20) had to identify either full words or 250 ms vowels isolated from the center of these words at nine f_os between 220 and 880 Hz. Listeners' identification performance was calculated with the bias free

non-parametric sensitivity measure A' from Signal Detection Theory (Stanislaw and Todorov, 1999). Results for A' were found extremely high in both conditions (i.e., words and isolated vowels) for each vowel at all investigated f_os. This means that vowels can remain identifiable at f_os exceeding the normal F_1-range of an intended vowel. In addition, the study also reveals that listeners do not rely on consonantal context phenomena such as formant transitions and co-articulation for their identification performance at high f_os. This indicates that vowel sounds may carry strong acoustic cues departing from common formant frequencies at high f_os.

Fig. 1–4 illustrates the methodological approach and shows the different fundamental frequencies that were tested in this study. It also gives an overview of the average F_1 values, which were obtained in a study on German vowels by Pätzold and Simpson (1997).

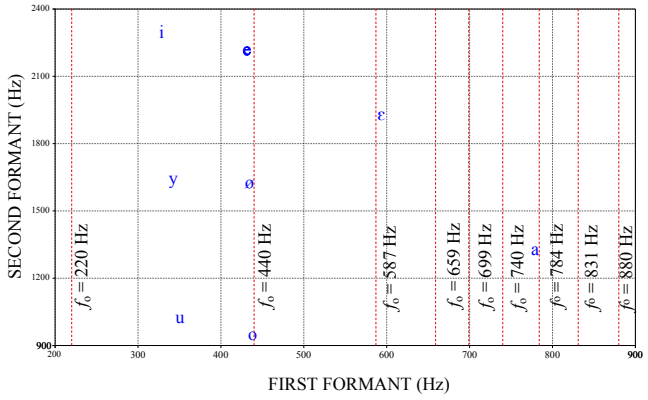

Figure 1–4: Schematic illustration of f_o exceeding the typical F_1 of German vowels. The figure shows an average F_1–F_2 space of the eight vowels investigated in this study. The dashed lines (red) mark the nine f_os at which the talker produced the vowels, and thus illustrate when an F_1 is approximately exceeded.

1.4 Study II

As listeners' identification performance for the eight vowels /i y e ø ɛ a o u/ was very good in a binary forced-choice task (study I), it was interesting to test whether the performance would remain equally good when a multiple-choice task is used (with all possible words as response options). Here, listeners' perception of the same German vowels produced by a female native German talker in words derived from a minimal pair with a single consonantal context (/'l–V–gən/) was assessed. The vowels were produced with steady spectral characteristics at nine f_os between 220 and 880 Hz. The vowel /u/ had to be excluded as "lugen" would have been the only meaningless word in German. A group of 28 German native listeners participated in an experiment, in which single words were randomly presented to the participants. The results show that vowel identification can be maintained with an error rate less than 20% up to an f_o of 740 Hz for /e ø ɛ/ and up to 880 Hz for /i y a o/. This confirms that vowel identification is possible in cases of f_o significantly exceeding F_1. In this study, also the role of neighboring vowels and vowel duration is analyzed and discussed in more detail.

1.5 Study III

In study III, the number of talkers was increased to three (all of them were female native German talkers) to test the influence of between-talker variability on listeners' identification performance. The same vowels already used in study I and II were produced in isolation, thus, not in a meaningful linguistic context. The fundamental frequency range was extended to eleven f_os between 220 and 1046 Hz, and a

multiple-choice identification task was used. In this closed-set identification task, 21 native German listeners were presented excised 700-ms vowel nuclei with quasi-flat f_o contours and resonance trajectories. The results differ from those of study I and II as they show that listeners can identify the point vowels /i a u/ at f_os up to almost 1 kHz, with a significant decrease for the vowels /y ɛ/ and a drop to chance level for the vowels /e ø o/ toward the upper f_os. Auditory excitation patterns reveal highly differentiable representations for /i a u/ that can be used as landmarks for vowel category perception at high f_os. This suggests that theories of vowel perception based on overall spectral shape will provide a fuller account of vowel perception than those based solely on formant frequency patterns.

1.6 Study IV

Paper IV is aimed at students and young researchers that are not yet familiar with the complex procedures of acoustic speech analysis, and its goal is to help them understand the basic principles of vowel analyses and to facilitate their choices when carrying out simple tasks such as pitch and formant analysis in *Praat* (Boersma and Weenink, 2016). The paper provides detailed information on the single steps involved in the standard procedures of the acoustic analyses of steady-state vowels. In addition, it highlights methodological problems, which go along with such analyses, and it offers practical insights that should help the reader to avoid them in their own projects. The study also introduces some intricate issues, for example, phenomena like *formant merging* and *spurious formants*), and it reviews problems that typically

go along with an extensive variation of f_o in large vowel databases.

1.7 Future work

The results presented in this cumulative thesis suggest that a theory of vowel perception based solely on formant frequency patterns cannot account for the relatively preserved performance listeners demonstrate in identifying vowels at high f_os. It seems likely that overall spectral shape features will play an important role in a coherent account of vowel perception generally. However, formal modeling of the relationship between the perceptual and physical spaces of vowels at high and low f_os is required for a convincing demonstration. In future studies, for example, it would be interesting to investigate the organization of the perceptual space at higher f_os by using multidimensional scaling analysis of cochlea-scaled spectra (see Friedrichs et al., 2016, for a first approach).

The results presented here could also be of interest in the context of the *quantal theory of speech* (Stevens, 1972, 1989, 2002). The identification rates and the differentiable auditory representations of /i a u/ at high f_os that are shown in study III are to the author's best knowledge the first empirical confirmation of the prediction made by Ken Stevens, who deduced from articulatory measurements that the point vowels should contain more robust features than all other vowels. This, for example, could lead to a better understanding of the fact that these categories are the only ones which can be found in almost all of the 6000+ documented languages of the world.

1.8 References

Adank, P., Smits, R., and Van Hout, R. (**2004**). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**(5), 3099–3107.

ANSI (**2013**). ANSI/ASA S1.1–2013, *Acoustical Terminology* , American National Standards Institute, Inc., New York.

Assmann, P. F., and Katz, W. F. (**2000**). "Time-varying spectral change in the vowels of children and adults," J. Acoust. Soc. Am. **108**(4), 1856–1866.

Benolken, M. S., and Swanson, C. E. (**1990**). "The effect of pitch-related changes on the perception of sung vowels," J. Acoust. Soc. Am. **87**(4), 1781–1785.

Bladon, A., & Fant, G. (**1978**). A two-formant model and the cardinal vowels. STL-QPSR, 19(1), 1–8.

Boersma, P., and Weenink, D. (**2016**)."Praat: Doing phonetics by computer [Computer program]" Version 6.0.15, retrieved March 23, 2016 from http://www.praat.org/ (Accessed July 30, 2016).

Chiba, T., and Kajiyama, M. (**1941**). "The vowel: Its nature and structure," Tokyo-Kaiseikan.

Deme, A. (**2014**). "Intelligibility of sung vowels: The effect of consonantal context and the onset of voicing," J. Voice **28**(4), 523–e19.

Diehl, R. L. (**2008**). "Acoustic and auditory phonetics: the adaptive design of speech sound systems," Philosophical Transactions of the Royal Society B: Biological Sciences **363**(1493), 965–978.

Fant, G. (**1960**). *Acoustic theory of Speech Production* (Mouton, The Hague, The Netherlands).

Friedrichs, D., Rosen, S., Iverson, P., Maurer, D., and Dellwo, V. (**2016**). "Mapping vowel categories at high fundamental frequencies using multidimensional scaling of cochlea-scaled spectra," J. Acoust. Soc. Am. **140**, 3219.

Gregg, J. W., and Scherer, R. C. (**2006**). "Vowel intelligibility in classical singing," J. Voice **20**(2), 198–210.

Hagiwara, R. (**1997**). "Dialect variation and formant frequency: The American English vowels revisited," J. Acoust. Soc. Am. **102**(1), 655–658.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**(6), 3013–3022.

Hollien H., Mendes-Schwartz A. P., and Nielsen K. (**2000**). "Perceptual confusions of high-pitched sung vowels," J. Voice **14**(2), 287–298.

Howie, J., and Delattre, P. (**1962**). "An experimental study of the effect of pitch on the intelligibility of vowels," Natl. Assoc. Teachers Singing Bull. **18**(4), 6–9.

Johnson, K. (**1997**). "Speech perception without speaker normalization: An exemplar model," in: K. Johnson, J. W. Mullennix (Ed.) *Talker variability in speech processing* (Academic Press, San Diego), pp. 145–165.

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). "Vocal tract resonances in singing: the soprano voice," J. Acoust. Soc. Am. **116**, 2434–2439.

Joos, M. (1948). "Acoustic phonetics," Language **24**(2), 5–136.

Lehiste, I., and Peterson, G. E. (**1959**). "Vowel Amplitude and Phonemic Stress in American English," J. Acoust. Soc. Am. **31**, 428–435.

Liberman, A. M., and Mattingly, I. G. (**1989**). "A specialization for speech perception," Science **243**(4890), 489–494.

Lindblom, B. (**1963**). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. **35**(11), 1773–1781.

Maurer, D., and Landis, T. (**1996**). "Intelligibility and spectral differences in high-pitched vowels," Folia Phoniatr. Logop. **48**(1), 1–10.

Maurer, D., d' Heureuse, C., and Landis, T. (**2000**). "Formant pattern ambiguity of vowel sounds," Int. J. Neurosci. **100**(1–4), 39–76.

Monsen, R. B., and Engebretson, A. M. (**1983**). "The accuracy of formant frequency measurements: A comparison of spectrographic analysis," J. Speech Hear. Res. **26**, 89–97.

Morozov, V. P. (**1965**). "Intelligibility in singing as a function of fundamental voice pitch," Soviet Physics–Acoustics **10**, 279–283.

Morrison, G. S., and Assmann, P. F. (Eds.). (**2012**). *Vowel inherent spectral change*. Springer Science & Business Media.

Nearey T., and Assmann P. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.

Pätzold, M., and Simpson, A. P. (**1997**). "Acoustic analysis of German vowels in the Kiel Corpus of Read Speech," Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel, 32, 215–247.

Peterson G. E., and Barney H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pierrehumbert, J. (**2001**). "Exemplar dynamics: Word frequency, lenition and contrast," in: Bybee, J. and Hopper, P. (Ed.) *Frequency and the emergence of linguistic structure* (John Benjamins Publishing Company, Amsterdam, The Netherlands), pp. 137–57.

Pols, L. C., Van der Kamp, L. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**(2B), 458–467.

Rakerd, B., and Verbrugge, R. R. (**1984**). "Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels," J. Acoust. Soc. Am. **77**(1), 296-301.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**(4497), 947–949.

Schnupp, J., Nelken, I., and King, A. (**2011**). *Auditory neuroscience: Making sense of sound* (The MIT Press, Cambridge, Massachusetts).

Scotto di Carlo, N., and Germain, A. (**1985**). "A perceptual study of the influence of pitch on the intelligibility of sung vowels," Phonetica **42**(4), 188–197.

Shepard, R. N. (**1972**). "Psychological representation of speech sounds," in: P. B. Denes, and E. E. David Jr. (Ed.) *Human Communication: A Unified View* (McGraw-Hill, New York), pp. 67–113.

Smith, L. A., and Scott, B. L. (**1980**). "Increasing the intelligibility of sung vowels," J. Acoust. Soc. Am. **67**, 1795–1797.

Stanislaw, H., and Todorov, N. (**1999**). "Calculation of signal detection theory measures," Behav. Res. Methods, Instrum., Comput. **31**, 137–149.

Stevens, K. N., and House, A. S. (**1955**). "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. **27**(3), 484–493.

Stevens, K. N. (**1972**). "The quantal nature of speech: Evidence from articulatory-acoustic data," in: P. B. Denes, and E. E. David Jr. (Ed.) *Human Communication: A Unified View* (McGraw-Hill, New York), pp. 51–66.

Stevens, K. N.(**1989**). "On the quantal nature of speech," J. Phon. **17**, 3–46.

Stevens, K. N. (**2000**). *Acoustic Phonetics*, Vol. 30 (The MIT Press, Cambridge Massachusetts).

Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**(4), 1872–1891.

Sundberg, J. (**2013**). "Perception of singing," in: D. Deutsch (Ed.), *Psychology of Music*, 3rd ed. (Academic Press, London, UK), pp. 69–106.

Sundberg, J., and Gauffin, J. (**1982**). "Amplitude of the voice source fundamental and the intelligibility of super pitch vowels," in: R. Carlson, and B. Granström (Ed.), *The representation of speech in the peripheral auditory system, proceedings of a symposium* (Elsevier Biomedical Press, Amsterdam, The Netherlands), pp. 223–228.

Terbeek, D., and Harshman, R. (**1972**). "Is Vowel Perception Non-Euclidean?," J. Acoust. Soc. Am. **51**(1A), 81.

Titze, I. R. (**2015**). "Balancing Odd and Even Harmonics in the Source Spectrum," J. of Singing **71**(3), 335.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S., and Wolfe, J. (**2015**). "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," J. Acoust. Soc. Am. **137**(5), 3005–3007.

Trainor, L. J., and Desjardins, R. N. (**2002**). "Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels," Psychonomic Bulletin & Review **9**(2), 335–340.

Yang, B. (**1996**). "A comparative study of American English and Korean vowels produced by male and female speakers," J. Phon. **24**(2), 245–261.

CHAPTER 2

Study I

The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz

The following article appeared in
The Journal of the Acoustical Society of America **138**(1), EL36–EL42,
and may be found at
D. Friedrichs, D. Maurer, and V. Dellwo, "The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz," J. Acoust. Soc. Am. **138**(1), EL36–EL42 (**2015**).

2.1 Abstract

In a between-subject perception task, listeners either identified full words or vowels isolated from these words at f_os between 220 and 880 Hz. They received two written words as response options (minimal pair with the stimulus vowel in contrastive position). Listeners' sensitivity (A') was extremely high in both conditions at all f_os, showing that the phonological function of vowels can also be maintained at high f_os. This indicates that vowel sounds may carry strong acoustic cues departing from common formant frequencies at high f_os and that listeners do not rely on consonantal context phenomena for their identification performance.

2.2 Introduction

Vocalic identification in naturally produced vowels at f_os exceeding the F_1 they typically reveal in citation-form words has so far mainly been a concern of singing research, particularly in Western classical singing ("legitimate style," henceforth: legit). By now there is a large body of evidence indicating that the identifiability of vowels decreases with increasing f_o. Identification of single vowels has been shown to be compromised when f_o significantly exceeds F_1; also referred to as oversinging Smith and Wolfe (2009). Early evidence for this position goes back to self-experiments by von Helmholtz (1885: p. 110) who found that the vowel /u/ loses its typical timbre from the musical note F3 ($\approx$ 175 Hz) upwards and shifts toward /o/. Howie and Delattre (1962) showed in an experiment with nine isolated sung vowels that listeners' identification performance decreased when f_o exceeded F_1. Hollien et al. (2000) showed for /i/, /a/, and /u/ that vowel category perception shifted mainly

to the one with the next higher F_1 as f_o increases (i.e., /i/ shifted to /ɪ/ then /e/ then /a/ when f_o exceeded F_1 for /i/, and /u/ shifted to /ʊ/, /o/, /ɔ/, /a/, respectively). Other studies have been more precise in identifying an absolute frequency at which the identification performance of listeners decreases. Sundberg (2012) provided evidence that this point corresponds to the musical note C5 ($\approx$ 523 Hz). Above this frequency, identification is heavily biased toward open vowels like /a/; from around 700 Hz it arrives at chance performance.

In legit, the communicative aim of producing intelligible utterances is typically in competition — and possibly secondary — to the aim of producing esthetical, sonorant, and powerful vocalizations. Legit singers, for example, adapt their resonance frequencies with the aim of enhancing their vocal power and homogeneity of timber, albeit at the expense of intelligibility (Joliveau et al., 2004). It is probably the result of this subordinate relevance of the communicative function of vowels in legit singing that vowel identification has primarily been studied in isolated vowels in the singing literature. From a linguistic point of view, however, the interest in vowels is typically in the functions they fulfill in speech communication like their phonological function in linguistic contrastive position (e.g., /e/ and /i/ may distinguish between the words desk and disc). Given the evidence for impoverished vowel identification performance of naturally produced vowels at high f_os above C5 from the singing literature, it seems conceivable that the phonological function of vowels in minimal pairs should also decrease with a substantial increase in f_o above this frequency. In the phonetic literature this question has so far not received much attention. Key studies on vocalic variability (Peterson and Barney, 1952; Hillenbrand et al., 1995;

Pätzold and Simpson, 1997) were primarily concerned with vowels at relatively low f_os (i.e., substantially below F_1 in citation-form words). This is also in line with observations that machine measurements of formants based on standard procedures (e.g., Linear Prediction Analysis) are highly problematic when $f_o >$ C5. It seems that it is implicitly taken for granted by phoneticians that the phonological function of vowels at high f_os should thus be poor. And this assumption seems justified as the probably strongest cues to vowel category identification–formant frequencies (in terms of determinable spectral maxima)–are poor when f_o increases significantly above C5.

Evidence exists which indicates that the consonantal environment of vowels at high f_os in real words enhances vowel identification. Smith and Scott (1980) reported higher identification rates for the front vowels /i/, /ɪ/, /ɛ/, /æ/ at f_os up to about 1100 Hz when they were produced in word consonant-vowel-consonant (CVC) context (/b/-V-/d/ resulting in bead, bid, bed, and bad) compared to the same vowels produced in isolation. One might assume that such results are driven by formant-transition phenomena between consonants and vowel (Strange et al., 1976), however, their impact on vowel identification has been strongly put into question (Diehl et al., 1981). It seems more likely that co-articulatory phenomena can explain the effect in Smith and Scott (1980), as the vocal tract configuration of a vowel is to a large degree in position during the surrounding consonants. This is particularly audible when one of the consonants is a voiceless fricative, characterized by a broadband noise source and ideally produced toward the rear end of the vocal tract (e.g., /heed/ and /hood/). In this case, listeners can likely profit from the acoustic characteristics of

the noise source shaped by the co-articulated vocal tract resonances of the vowel.

It also seems plausible that the poor identifiability of vowels at f_os higher than C5 is to a considerable degree the result of legit singing. This has already been suggested by Sundberg (2012) in particular, with reference to Smith and Scott (1980) who showed that in legit style, vowel intelligibility was poorer than in a condition in which singers raised their larynx and thus adapted their resonances to the increased f_o. Such evidence for a better identifiability of vowels at high f_os in a non-legit style was provided by Maurer et al. (2014) for a female singer of Cantonese opera. Listeners identification performance was drastically better than chance for 4 of her vowels (/i/, /a/, /o/, /u/) up to an f_o of 860 Hz. Because of the strong focus on voice esthetics in the singing literature, it remains unclear to what degree the phonological function is maintained at high f_os when a singer focuses on intelligibility rather than esthetics (i.e., when the singer does not sing in a specific singing style).

Here, we asked a trained female singer to produce minimal pairs including all long vowels of her native language (German) at varying f_o levels between 220 and 880 Hz focusing on the intelligibility of speech and, if necessary, ignoring esthetic qualities of her singing style. We extracted the steady state vocalic part (always 250 ms) of the word productions, resulting in two experimental conditions, words and isolated vowels. The fact that we made the singer produce the two words of each minimal pair in sequence, inevitably made her focus on the phonologically contrastive nature of the vowels during the production. In a between-subject design perception task, two groups of German native listeners identified the words extracted from the minimal pair productions being either presented as a full word stimulus (condition 1)

or an isolated vowel (condition 2). We extracted the words from pairs in which the difference in F_1 is expected to play a crucial role in the distinction of the vowels. This is true in particular, in minimal pairs contrasted by the front vowels /i/, /e/, /ø/, /y/, /e/, /a/ (15 possible pairs) and by the back vowels /u/, /o/ together with /a/ (3 possible pairs) in which between-category variability of F_2 is comparatively low but high for F_1. We tested to what degree listeners' ability to identify the correct word of a minimal pair decreased with increasing f_o for all vowel pairs. To avoid having varying numbers of response options and to test the words from the original pair productions, we provided listeners with binary response options (two words of the minimal pair). Should it hold that vowels with an $f_o >$ C5 lack acoustic category information then we would expect that: (i) For high-back vowels with low F_1 and low F_2, word identification performance should be poorest. (ii) Vowels in which f_o exceeds F_1 should more often be perceived as /a/-like, so for minimal pairs in which a contrast is built with the vowel /a/ listeners' identification performance should drop with higher f_o and listeners should be biased in their perception toward /a/. (iii) Should listeners rely on consonantal environment effects (co-articulation or formant transitions), it should be expected that identification performance drops drastically when such information is removed in vowels extracted from the carrier word (condition 2).

2.3 Methods

2.3.1 Subjects

Forty native German listeners without reported hearing impairments [20 male, 20 female; mean age = 26.78, standard deviation (s.d.) = 7.43], all students at the University of Zurich, participated in the experiment. Listeners were randomly divided into two groups (N = 20 per group; one group per condition [word and isolated vowel]; gender balanced across groups; mean age group 1: 29.75, s.d. = 8.73, group 2: 23.8, s.d. = 4.29).

2.3.2 Stimuli and apparatus

One female Musical Theatre singer (age 33; Swiss German native speaker, with excellent and trained pronunciation of Standard German) was recorded with a cardioid condenser microphone (Sennheiser MKH 40 P48 with pop shield, Wedemark-Wenne-bostel, Germany) on a PC via an audio interface (Fireface UCX, RME, Halmhausen, Germany) in a noise-controlled room at the University of Zurich. The singer was recorded in standing position; a drawn position reference on the floor helped the singer to keep a constant distance of about 30 cm to the microphone. The singer was selected based on her extended vocal range and a high skill of maintaining vowel quality at high f_os. The singer produced 18 German minimal pairs with a vocalic contrast in word mid position. All words were disyllabic and the contrasted vowels were part of the first syllable. Each contrastive vowel was in a CVC syllable. Mean duration of the vowels was 0.68 s (range: 0.58–1.11 s). Two sets of vowel contrasts were built, one with front vowels (/iː/, /yː/, /eː/, /øː/, /eː/,

/aː/) and one with back vowels together with /aː/ (/uː/, /oː/, /aː/). All vowels were contrasted with each other within the two different sets:

- Fifteen front vowel pairs: Biene-Bühne (/iː/-/yː/), siegen-Segen (/iː/-/eː/), biegen-Bögen (/iː/-/øː/), schielen-schälen (/iː/-/eː/), siegen-sagen (/iː/-/aː/), lügen-legen (/yː/-/eː/), rühren-Röhren (/yː/-/øː/), schürfen-schärfen (/yː/-/eː/), Sühne-Sahne (/yː/-/aː/), Lehne-Löhne (/eː/-/øː/), legen-lägen (/eː/-/eː/), Segen-sagen (/eː/-/aː/), töte-täte (/øː/-/eː/), Söhne-Sahne (/øː/-/aː/), schälen-Schalen (/eː/-/aː/).

- Three back vowel pairs (including /aː/): Buden-Boden (/uː/-/oː/), Buden-baden (/uː/-/aː/), Boden-baden (/oː/-/aː/)

The word pairs were recorded in two runs in AB and BA order. The singer was instructed to produce the minimal pairs as intelligible as possible. The word pair (AB or BA) that appeared to have the more perceptually salient vowel contrast to an investigator (second author) was chosen for the investigation. Each word pair was recorded at nine f_o levels (220, 440, 587, 659, 698, 740, 784, 831, 880 Hz) resulting in 162 minimal pairs (9 frequencies * 18 vowel contrasts). The lowest f_o level corresponded to the average f_o in citation-form words (Hillenbrand et al., 1995) and the entire frequency range of f_o produced was the range of the average F_1 for German vowels produced by women (Pätzold and Simpson, 1997). The respective

piano notes were presented as reference sounds to the singer via loudspeaker immediately preceding the production. f_o of the sound produced was measured in Praat (Boersma and Weenink, 2015) in the extracted vocalic parts. A maximum deviation from the reference f_o of 2.5% was found. Each of the two words from the chosen word pair recordings was extracted to serve as a stimulus in the word condition. For the isolated vowel condition, the steady state vowel centers were extracted with a duration of 250ms (125ms from the vowel mid point). At on- and offset the sounds were faded over 50ms by amplitude modulating the waveform with half a period of a cosine function [fade-in: $(1 - cos(x))/2$; fade-out: $(1 + cos(x))/2$]. Each stimulus was normalized for intensity (0 dB difference between stimuli); the overall output level was chosen by listeners individually.

2.3.3 Procedure

Two-word identification tests were carried out (one for each condition) in a small and noise controlled room using closed dynamic headphones (Beyerdynamic DT 770 Pro, 250 Ohm). In test 1, listeners were presented each word from each minimal pair (N = 324; 9 frequencies * 18 minimal pairs * 2 words) and saw a screen that contained 2 buttons (horizontally arranged) labeled with the words of the minimal pair (position—left/ right—was chosen randomly for each response option set). Above the response buttons the sentence Welches Wort hörst Du? (English: Which word do you hear?) could be read. Listener's task was thus to identify the word presented from the two response options (minimal pair) provided. Mm.1 contains an example of a word stimulus and Mm. 2 the respective isolated vowel stimulus derived from

this word.

- Mm. 1. Word stimulus “Buden” at 880 Hz; response options = “baden” and “Buden”. This is a file of type “wav” (118 Kb).

- Mm. 2. Isolated vowel stimulus /u:/ at 880 Hz extracted from the word Buden in Mm. 1; response options = ”baden” and “Buden”. This is a file of type “wav” (21 Kb).

After listeners made their choice they would hear the next stimulus automatically with a delay of 1 s. Listeners could not repeat a stimulus. Test 2 was identical to test 1 with the exception that an isolated vowel instead of a word was presented for identification. Above the response buttons, listeners could read the sentence Aus welchem Wort stammt der Vokal? (English: From which word did this vowel derive?). In test 2, listeners were explained that the presented vowel only referred to the contrasting vowel in the first syllable of the disyllabic word.

2.3.4 Data analysis

Listeners’ identification performance was calculated with the bias free nonparametric sensitivity measure A’ from Signal Detection Theory (Stanislaw and Todorov, 1999) with Praat scripts written by V. Dellwo according to formulas in Pallier (2002). One of the response options was arbitrarily assigned to the signal (signal vowel), the other to the noise (noise vowel). A “hit” was thus signal vowel

presented and responded, a "miss" was signal vowel presented but not responded, a "false alarm" was noise vowel presented but not responded, a "correct rejection" was noise vowel presented and responded. A' ranges between 0 and 1 with 0.5 being chance performance and 1 maximum performance. Values below 0.5 indicate response confusion. Listeners' response bias (i.e., a bias toward the vowel /aː/; see 2.3.2) was measured by B"$_D$ (Pallier, 2002). B"$_D$ ranges from 1 (maximum noise bias) to -1 (maximum signal bias). As each vowel was presented only once per listener, we pooled over listeners (N = 20) to calculate A' for each vowel pair at each f_o level and signal condition (N = 40; for example, the pair /iː/ vs. /eː/ was presented 20 times for /iː/ and 20 times for /eː/). So each A' value was calculated based on 40 responses by 20 listeners to a vowel pair.

2.4 Results

Figure 2–1 shows the distributions of A' at each f_o for the word and isolated vowel conditions of all minimal pairs; Fig. 2–2 shows the A' for word and isolated vowel conditions for each of the 18 minimal pairs separately. A' values for all investigated f_o levels (i.e., 220-880Hz) are high above chance level for both the word and isolated vowel conditions. For the word condition performance is at ceiling throughout all f_o levels. For the isolated vowel condition the interquartile range is roughly between A' = 0.9 and 1 at higher f_o levels. Two one-sample t-tests (one per condition; alpha = 0.01) testing the mean of the distribution against A' chance level (0.5) show that the effect was highly significant in both cases (words: t^{17} 83.43, $p < 0.001$; isolated vowels: $t^{17} = 29.23$, $p < 0.001$). The poorer performance for

isolated vowels in comparison to words was highly significant (Welch two-sample t-test: $t[222.75] = 7.32$, $p < 0.001$). To test that this effect could be replicated for individual f_o levels we carried out 18 one-sample t-tests, one for each f_o level (Bonferroni correction = 0.05/18 = 0.0028). T for 17 degrees of freedom ranged from 28.14 to 534.62. Each effect was highly significant ($p < 0.00028$).

To test the variation of A' between f_o levels we carried out a 9*2 two-factor analysis of variance (ANOVA) (f_o*condition). Results revealed a highly significant interaction ($F^{8,306} = 2.92$, $p < 0.005$), which was why we proceeded to calculate simple effects for each factor. Simple effects for f_o were studied by two one-factor ANOVAs (one for each condition). The effect for the word condition was not significant ($F^{8,153} = 1.01$, $p = 0.39$) and highly significant for the isolated vowel condition ($F^{8,153} = 5.14$, $p < 0.001$). This means that listeners had equally high performance in the word condition at all f_o levels and that performance decreased significantly with f_o in the isolated vowel condition. Simple effects for condition were tested by 9 two-sample t-tests (Welch) with a Bonferroni corrected alpha level of 0.0055 (0.05/9 f_o levels). A significant effect could be obtained for f_o level 4 (659 Hz) ($t[22.94] = 3.25$, $p < 0.005$) and a highly significant effect for level 9 (880 Hz) ($t[22.46] = 4.3$, $p < 0.0005$). It was surprising to obtain a significant effect at level 4 but not at the next higher levels (until level 9).

Listener bias calculation toward /aː/ (B''_D) is not meaningful when A' is high as it is only based on a small number of misses/false alarms (Stanislaw and Todorov, 1999). For this reason, we calculated B''_D only in case of the vowel pair /aː/-/eː/ under the isolated vowel condition for f_o of 831 and 880 Hz where A' values dropped to 0.81

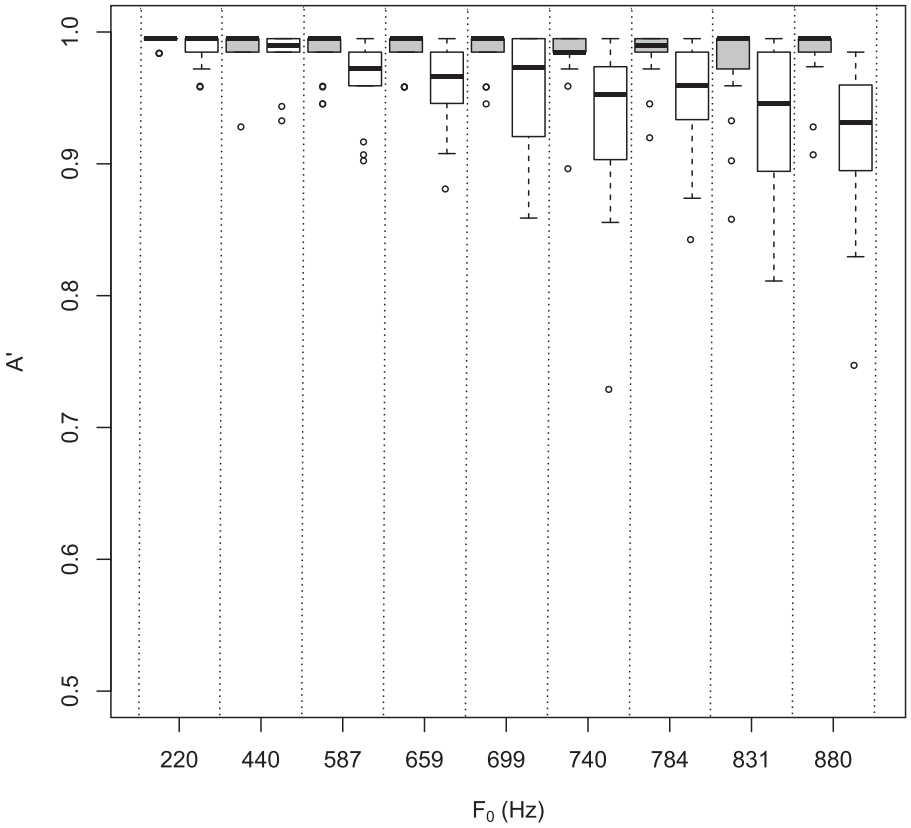

Figure 2–1: Box plots showing the distributions of A' (y axis) for all vowel pairs that were tested at nine f_o levels (x axis). Condition 1, words: white; condition 2, isolated vowels: gray. A' reaches from 0.5 (chance) to 1 (maximum performance).

and 0.75, respectively (Fig. 2–2).

We received B"$_D$ values of 0.8 and 0.89, respectively, indicating a strong signal bias (i.e., /aː/). This is small evidence for the hypothesis that under severe listening conditions (isolated vowels), listeners are biased in their perception of /eː/ toward /aː/ vowels at high f_os. However, this does not hold true for all other vowel contrasts

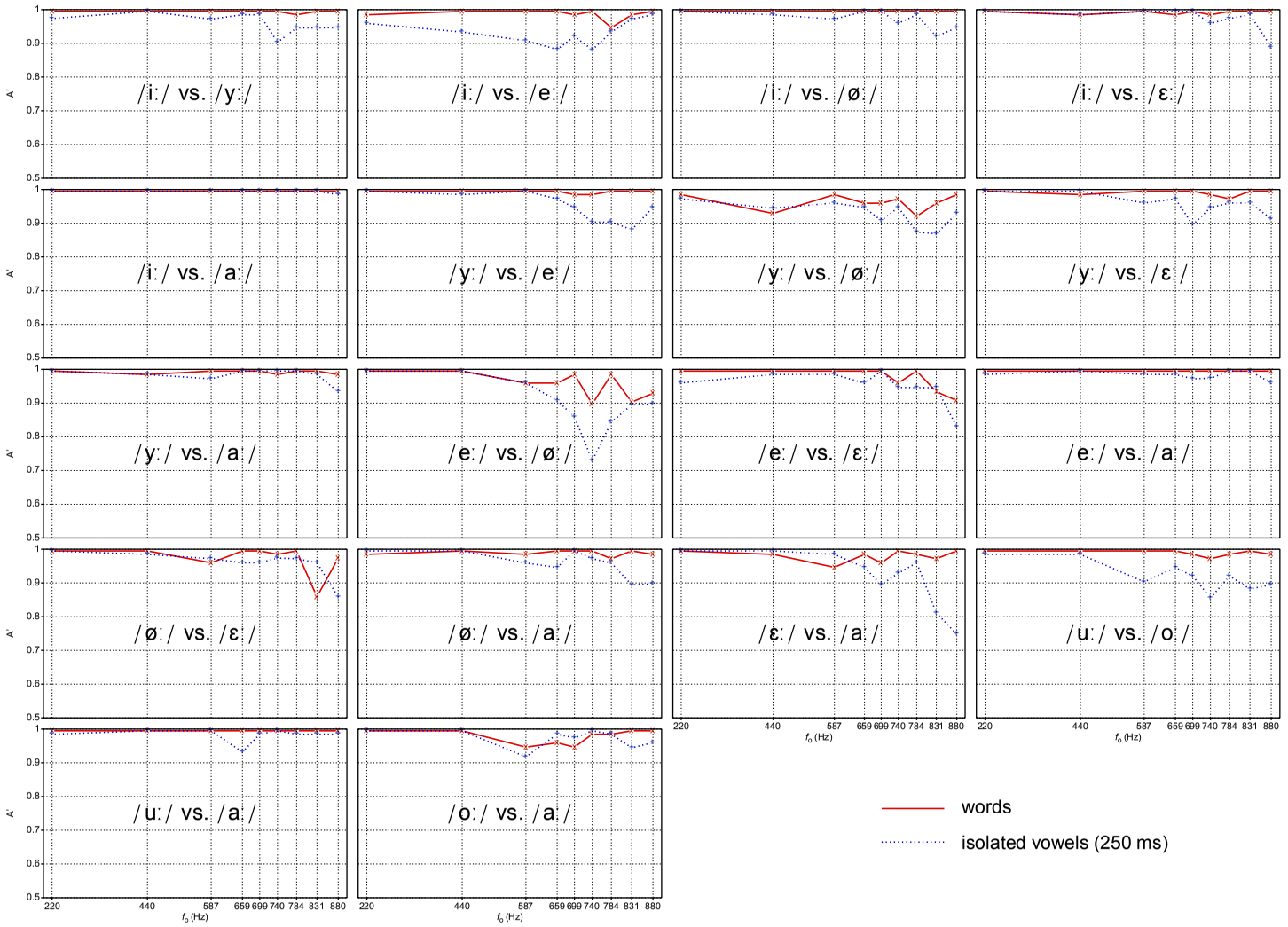

Figure 2–2: A' (y axis) for words (solid lines) and isolated vowels (dotted lines) for each of the minimal pair contrasts at the nine investigated f_o levels (x axis). A' reaches from 0.5 (chance level) to 1 (maximum performance).

tested that included /aː/ because the general performance for these vowel pairs was too high. For the high vowels together with /aː/ (/aː/-/iː/ and /aː/-/uː/), where the strongest decrease in performance should be expected because f_o exceeds F_1 drastically, the word identification performance was at ceiling level in both the word and the isolated vowel conditions.

Rare cases of higher A' for vowels tested in isolation compared to vowels tested in words could also be observed. This was true for /øː/-/aː/ at f_o = 220 Hz, /eː/-/aː/ at f_o = 440 and f_o = 587 Hz, /oː/-/aː/ at f_o = 659, f_o = 699, and f_o = 740 Hz, /iː/-/eː/ at f_o = 699 Hz, /yː/-/øː/ at f_o = 440 Hz, /yː/-/eː/ at f_o = 440 Hz, /eː/-/eː/ at f_o = 831 Hz, and /eː/-/øː/ at f_o = 587 and f_o = 831 Hz (Fig. 2–2). As these cases occurred non-systematically it seems likely that this was random variability or production variability in the data. It is unlikely that the speaker produced all vowel contrast equally well in each case.

2.5 Discussion

Results revealed that the phonological function of vowels can be surprisingly well maintained up to an f_o of at least 880 Hz. Even though an effect of signal condition (word vs. isolated vowels) was obtained, it must be concluded that the performance was extremely high under both conditions. The fact that the identification performance based on isolated vowels was only little below the performance of full word identification and always significantly above chance is support for the view that the isolated steady state part of the vowel contains sufficient vowel category information even at f_o = 880 Hz. It means that listeners do not rely on possible

co-articulatory or formant transition information in the surrounding consonants for their identification. What is the reason for this high identification performance in the isolated condition? It is possible that vowels produced in a linguistically meaningful environment contain clearer acoustic information to their category, in particular, when produced under severe conditions like at an f_o of 880 Hz. Isolated vowels which were produced in isolation by a speaker (Smith and Scott, 1980) resulted in lower identification results compared their context. It might also be the reason why Deme (2014) found no increase in performance of vowels in nonsense context environment in comparison to isolated vowels.

Listeners' ability to identify a word correctly in the word stimulus condition (condition 1) did not significantly decrease with increasing f_o up to 880 Hz for all vowel pairs tested. This was also true for the high back vowels for which we expected a strong decrease in performance. Therefore, we conclude that an increasing spectral under-sampling, which should inevitably lead to poorer vowel identification accuracy because of the sparser distribution of the harmonics, does not generally lead to a deterioration of the phonological function of vowels. In the case of isolated vowels, performance deteriorated significantly within a range of f_o from 220 toward 880 Hz. This might be weak evidence for a decrease in performance with the loss of consonantal context at vowels with $f_o >$ C5. It is also possible that the artificially generated fading at on- and offset in extracted vowels creates artifacts which contribute to this effect.

What role did F_1 and F_2 play for our results? It is unlikely that F_1 played a crucial role in vowel identification within a vowel pair concerning sounds at very different

levels of f_o. Words with high vowels containing maximally low F_1 and back vowels containing additionally maximally low F_2 (/iː/, /yː/, /eː/, /uː/, /øː/ and /oː/) could typically be identified at ceiling level across all f_o levels. We thus provided an example in which the phonological function of vowels is perfectly maintained when f_o substantially exceeds F_1. Concerning F_2, the pairs /uː/-/oː/ and /yː/-/øː/ in long German vowels are strongly under-sampled by f_o and 2 f_o when f_o = 880 Hz (see 2.3.2). In the case of /yː/-/øː/ the average F_2 frequencies in German are very close (1667 and 1646 Hz, respectively; Pätzold and Simpson, 1997). With 2 f_o at 1760 Hz it seems highly unlikely that F_2 was realized in a way in which it could contain subtle cues to vowel category in adjacent high back vowels. It thus seems unlikely that F_2 aided listeners in the word identification task in such cases. It is possible, however, that the position of vocal tract resonances between the harmonics influences the relative amplitude of higher harmonics, which may in return contain cues to vocalic category. To estimate the frequency of a vocal tract resonance by the harmonics, which amplitudes it is necessary for the listener to have experience with the spectrum of the vocal source. On the one hand it seems feasible that such knowledge was built up over the course of the experiment; on the other hand, we did not find any evidence that listeners performed less well for stimuli at 880 Hz when they incidentally occurred at the very beginning of the randomized stimulus set presentation. Future research will need to test whether listener's identification performance at f_o = 880 Hz improves with knowledge of a speaker's voice.

Listener bias toward /aː/ could typically not be tested in the minimal pairs containing this vowel as listeners' sensitivity was too high. The two cases, however, in

which the performance allowed measuring listener bias revealed that a bias toward /aː/ was present. Under more severe listening conditions or with more inexperienced speakers it seems conceivable that such an effect might occur more often.

Given the diverging results from previous studies, it is possible that individual speakers have a high impact on the results. Our speaker was a professional singer in Musical Theater style singing (i.e., non-legit) and is thus probably better suited to depart from legit's aesthetic resonance requirements. It thus seems feasible that our speaker was particularly well able to produce the vocalic contrastive information at high f_o levels due to extended vocal range, articulation, and professional training. Our example, however, proves that it is generally possible for speakers to produce vowels containing sufficient contrastive information at high f_os for reliable identification based on word presentations or isolated vowels. This finding is surprising, also for an individual speaker. It stands in contrast to the widely held view that cues to vowel category at f_os exceeding F_1 are technically impossible to produce. To generalize our findings, however, it will be important to study vowel recognition at high f_os with more speakers and possibly a larger variety of response options.

The finding now poses the question about which acoustic cues are responsible for the high word identification performance. Given that our speaker was able to produce contrasts between adjacent high vowel pairs (front as well as back pairs) which should be most affected by high f_os, it puts doubt on the widely held view that formant frequencies were the dominant cues in the word identification tasks. It is possible that other cues such as vowel inherent spectral change (Nearey and Assmann, 1986) explain the performance. The steady state parts in our vowels, however, did not

show typical spectral dynamic phenomena of continuous speech or isolated vowel productions. It thus seems questionable to what degree such phenomena might really explain listener identification performance in our vowels. Whichever cues future studies will reveal to be responsible for the result, it is possible that the cues to vowel identity at these high f_os might change our understanding of such cues at f_os typical for conversational speech (Maurer et al., 2000) and might thus contribute highly to our general understanding of human vowel perception.

2.6 Acknowledgements

We thank the professional singer, Heidy Suter, for producing the vowels for this study. This work was supported by the Swiss National Science Foundation (SNSF), Grant No. 100016_143943/1, and the Forschungskredit of the University of Zurich, Grant No. FK-14-062.

2.7 References and links

Boersma P., and Weenink D. (**2015**). "Praat: Doing phonetics by computer [Computer program]," Version 5.4.08, retrieved March 30, 2015 from http://www.praat.org/ (Last viewed March 30, 2015).

Deme A. (**2014**). "Intelligibility of sung vowels: The effect of consonantal context and the onset of voicing," J. Voice **28**, 523.e19–523.e25.

Diehl R. L., McCusker S. B., and Chapman L. S. (**1981**). "Perceiving vowels in isolation and in consonantal context," J. Acoust. Soc. Am. **69**(1), 239–248.

Hillenbrand J., Getty L. A., Clark M. J., and Wheeler K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111

Hollien H., Mendes-Schwartz A. P., and Nielsen K. (**2000**). "Perceptual confusions of high-pitched sung vowels," J. Voice **14**(2), 287–298.

Howie J., and Delattre P. (**1962**). "An experimental study of the effect of pitch on the intelligibility of vowels," NATS Bull. **18**, 6–9.

Joliveau E., Smith J., and Wolfe J. (**2004**). "Vocal tract resonances in singing: The soprano voice," J. Acoust. Soc. Am. **116**, 2434–2439.

Maurer D., D'Heureuse C., and Landis T. (**2000**). "Formant pattern ambiguity of vowel sounds," Int. J. Neurosci. **100**, 39–76.

Maurer D., Mok P., Friedrichs D., and Dellwo V. (**2014**). "Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer," in 15th Annual Conference of International Speech Communication Association, pp. 2132–2133.

Nearey T., and Assmann P. (**1986**). "Modeling the role of inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.

Pallier C. (**2002**). "Computing discriminability and bias with the R software," URL: http://www.pallier.org/ressources/aprime/aprime (Last viewed June 12, 2015).

Pätzold M., and Simpson A. (**1997**). "Acoustic analysis of German vowels in the Kiel Corpus of read speech," Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel **32**, 215–247.

Peterson G. E., and Barney H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Smith J., and Wolfe J. (**2009**). "Vowel-pitch matching in Wagner's operas: Implications for intelligibility and ease of singing," J. Acoust. Soc. Am. **125**, EL196–EL201.

Smith L. A., and Scott B. L. (**1980**). "Increasing the intelligibility of sung vowels," J. Acoust. Soc. Am. **67**, 1795–1797.

Stanislaw H., and Todorov N. (**1999**). "Calculation of signal detection theory measures," Behav. Res. Methods, Instrum., Comput. **31**, 137–149.

Strange W., Verbrugge R. R., Shankweiler D. P., and Edman T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Sundberg J. (**2012**). "Perception of singing," in Psychology of Music, 3rd ed., edited by D. Deutsch (Academic Press, London), pp. 69–106.

von Helmholtz H. (**1885**). On the Sensation of Tone (Dover, New York), republication 1954, 2nd ed. of the Ellis translation from 1885.

CHAPTER 3

Study II

Vowel identification at high fundamental frequencies in minimal pairs

The following article appeared in the
Proceedings of the 18th International Congress of Phonetic Sciences,
paper number 0434: 1–5, and may be found at
D. Friedrichs, D. Maurer, H. Suter, and V. Dellwo, "Vowel identification at high fundamental frequencies in minimal pairs," in: The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, UK, ISBN 978-0-85261-941-4, paper number 0434: 1–5 (**2015**).

3.1 Abstract

The question of vowel intelligibility as a function of f_o is still a matter of debate. Above all concerning vowel sounds produced at f_os exceeding vowel related statistical F_1 in citation-form words ('oversinging' F_1), it is unclear whether vowel category perception inevitably shifts towards the neighboring category with a higher F_1 or can be maintained in such cases. In this study, we tested listeners' perception of the long German vowels /i/, /y/, /e/, /ø/, /ɛ/, /a/, and /o/ produced by a trained female speaker in the context of minimal pair words (/l-V-gən/) at nine f_o-levels between 220 and 880 Hz. Results showed that vowel identification was maintained $< 80\%$ up to $f_o = 740$ Hz for /e/, /ø/, /ɛ/ and up to $f_o = 880$ Hz for /i/, /y/, /a/, and /o/. Thus, vowel identification could be maintained in cases of f_o significantly exceeding F_1. The role of neighboring vowels, vowel duration, and other productional and acoustical aspects relevant for vowel perception at different f_os is discussed.

3.2 Introduction

Several studies indicate that vowel intelligibility is compromised when the fundamental frequency (f_o) significantly exceeds the first formant frequency (F_1) in terms of both speaker-specific and statistical F_1 in citation-form words (the latter produced at $f_o \approx 220$ Hz). Early support for this view goes back to self-experiments by Helmholtz [8] who observed that the vowel /u/ shifts towards /o/ if the corresponding sound is produced at f_o exceeding 175 Hz. In a more detailed study, Howie and Delattre [11] investigated the intelligibility of the five English and four French

vowels /i/, /e/, /a/, /o/, /u/ and /y/, /ø/, /ɜ/, /ɛ/ sung by a baritone and a soprano (legit style) in isolation (hereafter V condition) at different levels of f_o (ranges of f_o = 132–396 Hz for the baritone, 264–1056 Hz for the soprano). They found that the identification performance of the listeners generally decreased when f_o exceeded F_1 of a vowel in question. Hollien et al. [10] studied the perception of the three corner vowels /a/, /i/, and /u/ produced in V condition by 18 professional male and female singers (legit and musical-theatre styles as well as singing teachers, ranges of f_o = 62–554 Hz for male singers, 165–1319 Hz for female singers). They found that when f_o of a sound exceeded F_1 of a back or front vowel, its perception shifted to the back or front vowel with the next higher F_1 and then to /a/, (i.e., /i/ shifted progressively to /ɪ/, /ɛ/ and then /a/, and /u/ shifted to /ʊ/, /o/, /ɔ/, and then /a/, respectively). Deme [4] investigated the perception of these three corner vowels produced by a single professional soprano singer (legit style) in V condition as well as in consonantal context, i.e., CVC condition. She found further support for this view for both production conditions. Identification rates dropped below 50% at f_o > 260 Hz for /i/ and > 350 Hz for /u/, while the identification rate of /a/ remained < 80% up to f_o = 988 Hz for unaltered V and CVC conditions, and < 60% up to the same f_o level for isolated vowels with the onset of voicing removed. In his attempt to define an upper limit for f_o of identifiable vowels in singing (legit style), however, Sundberg [22, 23] takes a more prudent stand. Searching for the highest percentage of correct identifications observed in various investigations of sung vowels [1, 7, 14, 16, 19, 20], he concluded for a possible identification > 80% of all vowels up to $f_o \approx$ 500 Hz although this frequency exceeds substantially F_1 of vowels such as /i/,

/y/, and /u/. As an explanation, he refers to pitch-dependent formant frequencies in singing, above all used by female singers, and states that, in such a singing technique, the decrease of vowel intelligibility is limited while loudness is gained [21, p. 129]. Moreover, referring to Smith and Scott [17], he indicates possible vowel identification for sounds at even higher levels of f_o, above all when produced in CVC condition [23, p.87], and referring to Gottfried and Chew [6], he points out the impact of a raised larynx for the production of intelligible vowel sounds. Smith and Scott [17] indeed reported results of a perceptual test of the front vowels /i/, /ɪ/, /ɛ/, and /æ/, produced by a soprano in legit style as well as with raised larynx, which showed an identification rate of 70% for all vowels up to $f_o = 880$ Hz in V condition with raised larynx and of 70–76% for all vowels up to $f_o = 1108$ Hz in CVC condition in legit style as well as with raised larynx. In a recent study of vowel perception in the singing and speaking in Cantonese Opera style, Maurer et al. [12] reported identification rates $> 80\%$ up to $f_o \approx 820$–860 Hz for the front and back vowels /i/, /a/, /ɔ/, and /u/ produced as syllables (C)V or (C)V:S. In line with this, yet concerning the perception of vowels at high f_os produced by untrained speakers, Maurer and Landis [13] reported high identification rates $> 90\%$ for all of the five long German vowels /i/, /e/, /a/, /o/, and /u/ produced by children in V condition up to $f_o \approx 660$ Hz, and for the corner vowels /i/, /a/, and /u/ up to $f_o \approx 840$ Hz. Thus, the results in the literature are inconsistent and we are left with the question whether or not vowel intelligibility is substantially compromised at f_os significantly exceeding typical F_1 values and, therefore, an increase of f_o is accompanied by perceptual shifts from vowels with low F_1 towards vowels with medium and high F_1. The present study

addresses this question by means of an investigation of the identifiability of the long German vowels /i/, /y/, /e/, /ø/, /ɛ/, /a/, /o/ produced by a female speaker (professional musical-theatre singer) in the context of minimal pair words ('/l–V–gən/) at nine levels of f_o in the range of 220–880 Hz and perceptually tested in a listening test involving 28 subjects. Hereafter, the vowels are separated into three subgroups, the front vowels /i/, /y/, /e/, /ø/, /ɛ/, and the vowel /a/, which was produced by the speaker within the range of /a-ɑ/ (no front-back classification applicable), and the back vowel /o/. The vowel /u/ has not been included because the word lugen is not a commonly known and used lexical unit in the German language.

3.3 Methods

3.3.1 Subjects

A group of 28 Swiss German native listeners (all students at the University of Zurich; 15 female, 13 male; mean age = 23.1, sd = 1.5) participated in the experiment. None of them reported any kind of hearing impairments.

3.3.2 Stimuli and apparatus

A female speaker (age = 33; Swiss German native speaker, professional musical-theatre singer) produced the German vowels /i/, /y/, /e/, /ø/, /ɛ/, /a/, /o/ in /'l–V–gən/ context at f_o of 220, 440, 587, 659, 699, 740, 784, 831 and 880 Hz. Digital recordings (44.1 kHz sampling rate, 24-bit resolution) were made in a noise-controlled room at the University of Zurich using a cardioid condenser microphone (Sennheiser MKH 40 P48 with pop shield) and an audio interface (Fireface UCX)

connected to a PC. The speaker-microphone distance was 30 cm. For each of the f_os investigated, the speaker was instructed to produce the vowels in word pairs as minimal pairs within two sets of vowel contrasts, front vowels and /a/, and the back vowel /o/ and /a/, in AB and BA order. Thus, all vowels were contrasted with each other within the sets of /i/, /y/, /e/, /ø/, /a/ and /o/, /a/ in the two possible orders of the words in a pair, e.g., liegen vs. lügen, lügen vs. liegen, liegen vs. legen, legen vs. liegen etc. Piano notes were presented as reference sounds to the speaker via loudspeaker immediately preceding the production. Listening to the utterances (first and second author), for each vowel and each level of f_o, the word token that appeared to manifest the optimal correspondence between the intended and the perceived vowel category was chosen for further investigation. Thus, for each level of f_o, each of the seven vowels was represented by one /'l–V–gən/ token(N = 63; 7 words * 9 f_os). Mean f_o was calculated for 250 ms in the middle of a vowel sound in Praat [3] using the algorithm described in [2]. A maximum deviation from the intended f_o of 1.9% was found.

3.3.3 Procedure (listening test)

Single stimuli /'l–V–gən/ were randomly presented to the participants of the listening test via closed dynamic headphones (Beyerdynamic DT 770 Pro) in a small and noise-controlled room. On a computer screen, buttons labeled with the seven investigated words were randomly arranged in a circle to account for a potential directional bias of the listeners. Above the response buttons, the sentence *Welches Wort hörst Du?* (Which word do you hear?) could be read. When listening to

a word, subjects were asked to assign one of the words presented on the screen (seven-alternative forced choice word identification task). After a response, the next stimulus was presented with a delay of 1 sec.

3.3.4 Data analysis

To approximate the speaker-specific F_1 at a level of f_o comparable to statistical F_1 in citation-form words, mean F_1 values were calculated for the steady-state mid 250 ms of the vowels produced at $f_o = 220$ Hz, using Praat (Burg algorithm for LPC, default settings for female speakers). Calculated formant frequency values were double-checked on the basis of the respective spectrograms. The total duration of all vowel sounds from onset to offset was measured in Praat with the help of wideband spectrograms. Durations were averaged for each vowel category to investigate the influence of durational information on vowel identification at higher f_os. Identification rates (hereafter ID rate) in terms of the correspondences between the intended and the perceived vowels were determined for each f_o. Referring to Sundberg [23, p. 87], an ID rate $< 80\%$ was considered as accurate vowel intelligibility, and only the cases with a lower rate were investigated in more detail.

3.4 Results

Table 3–1 shows a comparison of the statistical mean F_1 in citation-form words ($F_{1[}stat$]) obtained by Pätzold and Simpson [15] and the mean F_1 estimations for each vowel produced by the investigated speaker in /'l–V–gən/ context at $f_o = 220$ Hz. The values for /i/, /e/, /ø/, /o/, /a/ $F_{1[}speaker$] are in good accordance.

However, $F_{1[}speaker]$ is substantially lower than $F_{1[}stat]$ for /y/. Since Pätzold and Simpson [15] do not report values for the long vowel /ɛ/, and because no reliable estimation of $F_{1[}speaker]$ in terms of a correspondence of LPC values and spectrogram for /ɛ/ was possible, no corresponding F_1 were considered for this vowel.

Vowel	$F_{1[stat]}$ (Hz)	$F_{1[speaker]}$ (Hz)
i:	329	367
y:	342	240
e:	431	442
ø:	434	421
o:	438	416
ɛ:	--	--
a:	779	865

Table 3–1: Mean statistical F_1 values ($F_{1(}stat)$) for Standard German vowels and mean F_1 estimations (mid 250 ms) for the vowels produced by the female speaker at $f_o = 220$ Hz ($F_{1(}speaker)$).

The listeners' identification performance is shown in the confusion matrices in Table 3–2 (one matrix for each f_o level). With the exception of /ø/ at $f_o = 587$ Hz, all ID rates proved to be $> 23/28$, i.e., $< 80\%$ up to f_o of 740 Hz. For the vowels /i/, /y/, /o/, and /a/ this even holds true up to f_o of 880 Hz. In addition to the case of /ø/ at $f_o = 587$ Hz, substantial confusions (ID rates $< 80\%$) and corresponding perceptual shifts towards other vowel categories occur for the utterances of the vowels /e/, /ø/, and /ɛ/ in the f_o range of 740–880 Hz (see Table 3–3).

F_0 = 220 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	28	0	0	0	0	0	28
e:	0	0	25	0	0	3	0	28
ø:	0	0	0	28	0	0	0	28
o:	1	1	0	0	26	0	0	28
ɛ:	0	0	0	0	0	27	1	28
a:	0	1	0	0	0	0	27	28
	29	30	25	28	26	30	28	196

F_0 = 440 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	27	0	1	0	0	0	0	28
y:	0	28	0	0	0	0	0	28
e:	0	0	26	1	0	1	0	28
ø:	0	0	0	28	0	0	0	28
o:	0	0	0	0	27	0	1	28
ɛ:	0	0	0	0	0	28	0	28
a:	0	0	0	0	0	0	28	28
	27	28	27	29	27	29	29	196

F_0 = 587 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	27	0	0	1	0	0	28
e:	0	0	26	0	0	2	0	28
ø:	0	1	1	10	0	15	1	28
o:	0	0	1	0	27	0	0	28
ɛ:	0	0	0	0	0	28	0	28
a:	0	0	0	0	3	0	25	28
	28	28	28	10	31	45	26	196

F_0 = 659 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	28	0	0	0	0	0	28
e:	0	0	24	1	0	2	1	28
ø:	0	3	1	24	0	0	0	28
o:	1	0	0	0	27	0	0	28
ɛ:	0	0	0	0	0	27	1	28
a:	0	0	0	0	4	0	24	28
	29	31	25	25	31	29	26	196

F_0 = 699 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	27	0	0	0	1	0	28
e:	1	0	23	0	0	4	0	28
ø:	0	0	0	28	0	0	0	28
o:	0	0	1	0	27	0	0	28
ɛ:	0	0	0	0	0	28	0	28
a:	0	0	0	0	3	0	25	28
	29	27	24	28	30	33	25	196

F_0 = 740 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	26	0	0	0	0	2	0	28
y:	0	28	0	0	0	0	0	28
e:	0	2	23	1	0	2	0	28
ø:	0	2	0	22	3	1	0	28
o:	0	0	0	0	28	0	0	28
ɛ:	0	0	1	1	0	23	3	28
a:	0	0	1	0	1	0	26	28
	26	32	25	24	32	28	29	196

F_0 = 784 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	1	26	1	0	0	0	0	28
e:	4	0	15	4	0	5	0	28
ø:	0	1	1	26	0	0	0	28
o:	0	0	0	0	28	0	0	28
ɛ:	0	0	0	0	0	16	12	28
a:	0	0	0	0	1	1	26	28
	33	27	17	30	29	22	38	196

F_0 = 831 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	28	0	0	0	0	0	28
e:	0	0	17	0	0	11	0	28
ø:	0	0	9	6	0	12	1	28
o:	0	1	0	0	27	0	0	28
ɛ:	0	0	0	0	1	16	11	28
a:	0	0	0	0	1	0	27	28
	28	29	26	6	29	39	39	196

F_0 = 880 Hz

	i:	y:	e:	ø:	o:	ɛ:	a:	
i:	28	0	0	0	0	0	0	28
y:	0	28	0	0	0	0	0	28
e:	0	0	14	2	1	11	0	28
ø:	0	2	1	18	6	1	0	28
o:	0	0	0	0	28	0	0	28
ɛ:	0	0	0	0	0	24	4	28
a:	0	0	0	0	0	1	27	28
	28	30	15	20	35	37	31	196

Table 3–2: Confusion matrices showing intended vowels (column 1) versus perceived vowels (column 2–8) for all f_os investigated. Number of listeners = 28. In a single matrix, the bottom row shows the total number of vowel category responses.

For /e/, ID rate was $< 80\%$ for f_o of 784, 831 and 880 Hz. However, a strong and oriented perceptual shift was only found for f_o of 831 and 880 Hz in terms of a shift towards /ɛ/. This vowel can be considered as related to the vowel with the next higher F_1 (see, e.g., [9]). For /ø/, ID rate was $< 80\%$ for f_o of 587, 740, 831 and

880 Hz. Strong and oriented shifts were only found for f_o of 587 and 831 Hz towards /ɛ/ and /e/, respectively. /ɛ/ can again be considered as related to the vowel with the next higher F_1. However, this is not the case for /e/, for which the vowel-related differences in the formant patterns at f_o in citation form words concern F_2 and F_3 (see [15]). For /ɛ/, ID rate was $< 80\%$ for f_o of 784 and 831 Hz. Strong and oriented shifts were found for both levels of f_o towards /a/, i.e., to the vowel with the highest F_1.

V_{int}	F_0	i:	y:	e:	ø:	o:	ɛ:	a:	oriented confusion
ø:	587	0	1	1	**10**	0	15	1	ø:-ɛ:
ø:	740	0	2	0	**22**	3	1	0	
e:	784	4	0	**15**	4	0	5	0	
ɛ:	784	0	0	0	0	0	**16**	12	ɛ:-a:
e:	831	0	0	**17**	0	0	11	0	e:-ɛ:
ø:	831	0	0	9	**6**	0	12	1	ø:-ɛ:,ø:-e:
ɛ:	831	0	0	0	0	1	**16**	11	ɛ:-a:
e:	880	0	0	**14**	2	1	11	0	e:-ɛ:
ø:	880	0	2	1	**18**	6	1	0	

Table 3–3: Intended vowels at the levels of f_o for which ID rates dropped below 80% ($< 23/28$ correct responses). Strong oriented confusions (perceptual vowel category shifts $> 50\%$ of the number of correct responses) are displayed on the right.

Mean duration of the vowel sounds was 622 ms (sd = 99 ms; range = 430–868 ms). One-way ANOVA revealed significant difference ($F^{(6,56)} = 2.34$, $p < .05$) in sound duration of the seven vowels investigated. Tukey's HSD tests only revealed a significant difference in sound duration for /y/ and /a/ ($p = .04$), and /y/ and /o/ ($p = .03$). No significant difference could be found in sound duration for all vowel

pairs of /i/, /o/, /e/, /ø/, /ɛ/, and /a/ ($p < .88$). Figure 3–1 shows the distribution of the sound duration for the investigated vowels.

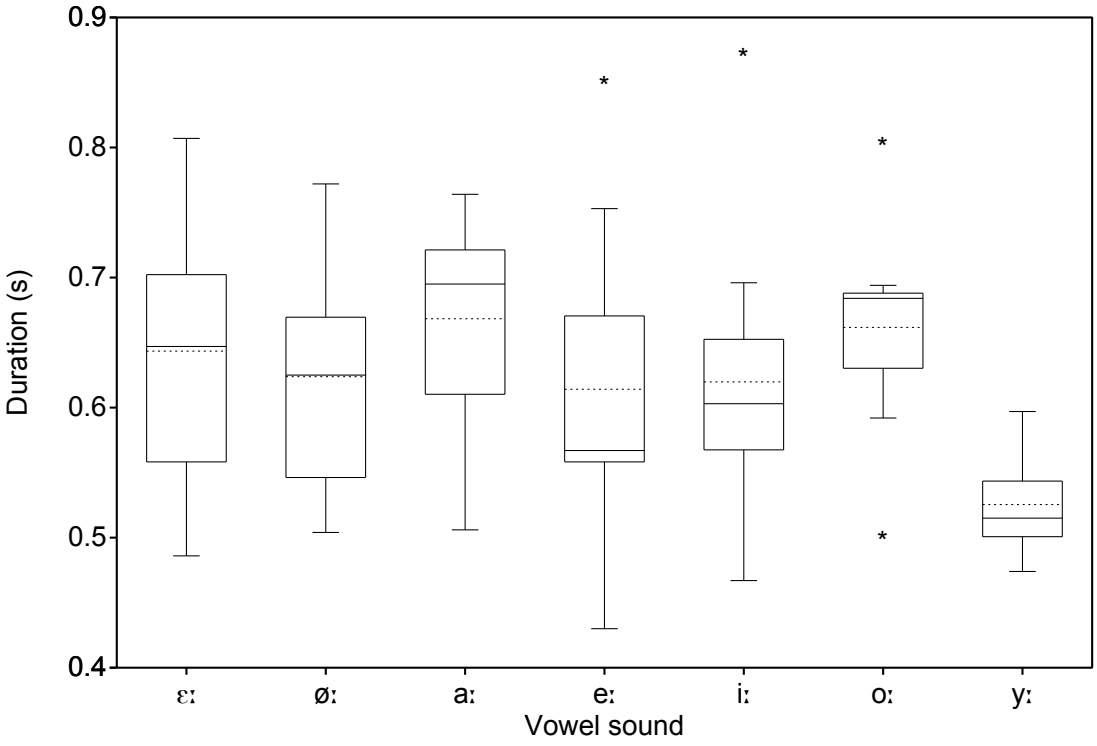

Figure 3–1: Boxplots representing the distribution of the duration of the vowel sounds.

3.5 Discussion

The vowels /i/, /y/, /o/, and /a/ were consistently identified with ID rates $< 80\%$ up to f_o of 880 Hz. Since F_1 of the speaker at f_o of 220 Hz corresponds well with statistical F_1 in citation-form words, since /i/ and /y/ are related to the lowest levels of corresponding F_1, and since f_o of 880 Hz corresponds to the highest F_1 values for

all vowels investigated (see Table 3–1), the results indicate that 'oversinging' statistical F_1 does neither inevitably compromise vowel perception, nor does the perceived vowel category inevitably shift to the category with the next higher F_1. Contrarily, a consistent vowel perception can be maintained independent of statistical F_1. With the exception of the case of /ø/ at $f_o = 587$ Hz (possibly due to production inconsistency), the finding that all vowels were identified with a rate $< 80\%$ up to f_o of 740 Hz strongly supports such a conclusion. Concerning the decrease of the identification rates for /e/, /ø/, and /ɛ/ above all for $f_o > 740$ Hz, a tendency of a shift in the perceived vowel category to the one with the next higher F_1 is indicated by the results. However, the tendency is inconsistent, i.e., it was not found for all combinations of vowels and f_o-level > 740 Hz. Moreover, an alternative interpretation also has to be considered. Discussing possible confusions in terms of shifts towards non-intended vowel categories may have to account for the entire formant patterns of the vowels under investigation and the respective 'density' of neighboring vowel categories according to their placement in the vowel quadrilateral. This would explain why, in the present study, (i) corner vowels were identified more correctly than non-corner vowels, (ii) no strong confusion was found for /o/ (the only back vowel) but some pronounced confusions were found for /e/, /ø/, and /ɛ/ (three of five front vowels; note also that F_1 is comparable for /o/, /e/, and /ø/), (iii) identification of /i/, /y/ (closed front vowels) proved to be more consistent than of /e/, /ø/ (closed-mid front vowels) and of /ɛ/ (open-mid front vowel); (iv) perceptual shifts were not limited towards the vowel category with a higher F_1 (see Table 3–3, vowels and f_o without clear shift tendencies, and the shift /ø/-/e/ at f_o of 831 Hz). Statistical

analysis does not indicate a clear relation between sound duration and identification performance of the vowels. In line with earlier studies of possible vowel identification at high pitches, the present investigation again shows that vowel perception at very different levels of f_o cannot be directly related to vowel-specific formant patterns as given for citation-form words. Although a high vocal ability of the speaker and a modified vowel production (e.g., raised larynx, adaption of articulation) may play a crucial role for the present findings, and in addition, dynamic spectral characteristics because of the consonantal context (see e.g., [17, 18]; however, for controversial position, see [5]), and meaning of the /'l-V-gən/ tokens may also have a substantial impact on vowel perception, these factors do not allow for a satisfactory explanation concerning the acoustic cues listeners referred to when perceiving the vowels at the very different levels of f_o. Thus, the acoustic cues of vowel perception including all f_o of vowel identifiability are still a matter of investigation for future research.

3.6 Acknowledgements

This work was supported by the Forschungskredit of the University of Zurich, grant FK-14-062, and the Swiss National Science Foundation (SNSF), grant 100016_143943/1.

3.7 References

[1] Benolken, M.S., Swanson, C.E. (**1990**). The effect of pitch-related changes on the perception of sung vowels. J. Acoust. Soc. Am. **87**(4), 1781–1785.

[2] Boersma, P. (**1993**). Accurate short-term analysis of the fundamenta frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences **17**, 97–110.

[3] Boersma, P., Weenink, D. (**2014**). Praat: doing phonetics by computer [Computer program]. Version 5.4.04, retrieved 29 December 2014 from http://www.praat.org/

[4] Deme, A. (**2014**). Intelligibility of sung vowels: the effect of consonantal context and the onset of voicing. J. Voice **28**, 523.e19–25.

[5] Diehl, R. L., McCusker, S. B., Chapman, L. S. (**1981**). Perceiving vowels in isolation and in consonantal context. J. Acoust. Soc. Am. **69**(1), 239–248.

[6] Gottfried, T.L., Chew, S.L. (**1986**). Intelligibility of vowels sung by a countertenor. J. Acoust. Soc. Am. **79**(1), 124–130.

[7] Gregg, J.W., Scherer, R.C. (**2006**). Vowel intelligibility in classical singing. J. Voice **20**, 198–210.

[8] Helmholtz, H. von, (**1870**). Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. Braunschweig: Vieweg.

[9] Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K. (**1995**). Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**(5), 3099–3111.

[10] Hollien, H., Mendes-Schwartz, A.P., Nielsen, K. (**2000**). Perceptual confusions of high-pitched sung vowels. J. Voice **14**, 287–298.

[11] Howie, J., Delattre, P. (**1962**). An experimental study of the effect of pitch on the intelligibility of vowels. Natl. Assoc. Teach. Sing., 385–394.

[12] Maurer, D., Mok, P., Friedrichs, D., Dellwo, V. (**2014**). Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer. Fifteenth Annu. Conf. Int. Speech Commun. Assoc. Singapore, 2132–2133.

[13] Maurer, D., Landis, T. (**1996**). Intelligibility and spectral differences in high pitched vowels. Folia Phoniatr. Logop. **48**, 1–10.

[14] Morozov, V.P. (**1965**). Intelligibility in Singing as a Function of Fundamental Voice Pitch. Sov. Physics–Acoustics (Translated from Akust. Zhurnal, **Vol. 10**, No. 3, pp. 330–334, July–September, 1964), 395–402.

[15] Pätzold, M., Simpson, A. (**1997**). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel **32**, 215–247.

[16] Scotto Di Carlo, N., Germain, A. (**1985**). A perceptual study of the influence of pitch on the intelligibility of sung vowels. Phonetica **42**, 188–197.

[17] Smith, L.A., Scott, B.L. (**1980**). Increasing the intelligibility of sung vowels. J. Acoust. Soc. Am. **67**(5), 1795–1797.

[18] Strange, W., Verbrugge, R.R., Shankweiler, D.P., Edman, T.R. (**1976**). Consonant environment specifies vowel identity. J. Acoust. Soc. Am. **60**(1), 213–224.

[19] Sundberg, J. (**1977**). Vibrato and vowel identification. Arch. Acoust., 257–266.

[20] Sundberg, J. Gauffin, J. (**1982**). Amplitude of the voice source fundamental and the intelligibility of super pitch vowels. In: Carlson R., Granström, B. (eds),

The representation of speech in the peripheral auditory system. Amsterdam: Elsevier Biomedlical Press, 223–228.

[21] Sundberg, J. (**1987**). The Science of the Singing Voice. DeKalb, IL: Northern Illinois University Press.

[22] Sundberg, J. (**1994**). Perceptual aspects of singing. J. Voice **8**, 106–122.

[23] Sundberg, J. (**2012**). Perception of Singing. In: Deutsch, D. (eds), The Psychology of Music. London: Academic Press, 69–106.

CHAPTER 4

Study III

Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks

The following article appeared in

The Journal of the Acoustical Society of America **142**(2), 1025–1033,

and may be found at

D. Friedrichs, D. Maurer, S. Rosen, and V. Dellwo, "Vowel recognition at fundamental frequencies up to 1 kHz reveals point vowels as acoustic landmarks," J. Acoust. Soc. Am. **142**(2), 1025–1033 (**2017**).

4.1 Abstract

The phonological function of vowels can be maintained at fundamental frequencies (f_o) up to 880 Hz [Friedrichs, Maurer, and Dellwo (**2015**). J. Acoust. Soc. Am. **138** , EL36–EL42]. Here, the influence of talker variability and multiple response options on vowel recognition at high f_os is assessed. The stimuli (n=264) consisted of eight isolated vowels (/i y e ø ɛ a o u/) produced by three female native German talkers at eleven f_os within a range of 220–1046 Hz. In a closed-set identification task, 21 listeners were presented excised 700-ms vowel nuclei with quasi-flat f_o contours and resonance trajectories. The results show that listeners can identify the point vowels /i a u/ at f_os up to almost 1 kHz, with a significant decrease for the vowels /y ɛ/ and a drop to chance level for the vowels /e ø o/ towards the upper f_os. Auditory excitation patterns reveal highly differentiable representations for /i a u/ that can be used as landmarks for vowel category perception at high f_os. These results suggest that theories of vowel perception based on overall spectral shape will provide a fuller account of vowel perception than those based solely on formant frequency patterns.

4.2 Introduction

Patterns of formant frequencies are commonly assumed to be the most salient cues to vowel perception. The assumption that the vowel identification process is mainly driven by such an underlying acoustic representation contributes largely to the pervasive idea that listeners' ability to recognize vowels has to be poor at very high fundamental frequencies (f_o) due to a sparse sampling of the vocal tract transfer

function. This holds true, in particular, when the normal range of the first formant frequency (F_1) is exceeded by f_o, and the higher formants are poorly specified due to a wide spacing of the harmonics.

Support for this view is mainly provided by studies on Western operatic singing. Howie and Delattre (1962), for example, found in a study on the perception of high-pitched vowels (f_o range 132–1056 Hz) sung by a baritone and a soprano that vowels lose their identity increasingly with increasing f_o. This degradation starts with the categories usually characterized by a low F_1 (i.e., high vowels such as /i/ and /u/) and leaving only those with the highest F_1 (i.e., low vowels such as /a/ and /ɑ/) identifiable at very high f_os. Ever since, numerous studies have reported that only /a/-like vowels can remain identifiable at the highest musical notes near 1 kHz (see Sundberg, 2013, p. 87, for an overview). It seems plausible, however, that this loss of vowel contrast is primarily due to articulatory changes applied by Western operatic singers when they perform at higher pitches. In experimental studies such as Joliveau et al. (2004) it has been shown, for example, that sopranos shift the first resonant frequency (f_{R1}) of their vocal tract – and thus F_1 – to the vicinity of f_o as soon as f_o drastically exceeds the normal range of f_{R1} of an intended vowel. This tuning of f_{R1} is achieved by increasing the jaw opening and reducing the maximum constriction of the vocal tract (Sundberg, 1975; Sundberg, 2013). As f_o gains considerable amplitude when being closer to a resonant frequency, these maneuvers may help a singer to maintain vocal power and timbral homogeneity (Smith and Wolfe, 2009). However, the acoustic modifications associated with shifting a resonant frequency may lead to ambiguous formant frequency patterns and consequently to a confusion

of vowel categories.

Given this situation, it is surprising that few studies have investigated vowel recognition outside Western operatic singing at very high f_os as there is evidence that even a sparsely sampled vocal tract transfer function still carries information, which can be used by listeners to recognize different vowels, despite a likely absence of the supposed F_1 and an undersampling of the higher formants. Smith and Scott (1980), for example, reported listeners' identification performance significantly above chance level (mean of 70% correct) for the four front vowels /i ɪ ɛ æ/, which were produced by a soprano in isolation at an f_o of about 880 Hz (i.e., the musical note A5) with a raised larynx (i.e., a shortened vocal tract), and thus not in an articulation mode typical for Western operatic singers. When asked to produce the same vowels in her operatic singing style, identification dropped to a mean of 4% correct at the same f_o. Maurer and Landis (1996) showed that infant and adult talkers can produce identifiable versions of the vowels /i a o u/ but not of /e/ at an f_o between about 500–870 Hz that was individually chosen by the talker. In a more recent study, Maurer et al. (2014) investigated the high-pitched vowels /i y œ a ɔ u/ produced by a female Cantonese opera singer in isolation and monosyllabic consonant-vowel utterances and found that /i a ɔ u/ could be identified by more than 80% of the listeners within an f_o range of 820–860 Hz. In a study using a two-alternative forced choice task, Friedrichs et al. (2015a) provided evidence that the phonological function of the eight vowels /i y e ø ɛ a o u/ (i.e., the function they fulfil in linguistic contrastive position to help listeners distinguish between words) can be maintained at f_os up to at least 880 Hz when they were produced in minimal pairs. These judgments were

made on excised steady-state vowel nuclei (250 ms) excluding consonantal context phenomena such as co-articulation and formant transitions. This is particularly surprising for vowels that typically have a low F_1 that were tested in combination with adjacent vowels with similar F_2 (e.g., /i/ vs. /e/ and /u/ vs. /o/), because an absent F_1 has been argued to make vowels with a similar F_2 indistinguishable (Smith and Wolfe, 2009, p. E196; see Ito et al., 2001, for contradictory results). In a follow-up study (Friedrichs et al., 2015b), a female talker produced the same vowels except /u/ in the German word context /'l–V–gən/ (/u/ was excluded as it would have resulted in a meaningless utterance), and a multiple-choice identification task was used. It was found that the words including /i y a o/ remained identifiable – and thus the vowels' phonological function could be maintained – throughout the investigated f_o range from 220 to 880 Hz. For the vowels /e ø ɛ/, however, a significant decrease was observed in listeners' identification performance within this range (for /ø/ from about 587 Hz and for /e ɛ/ from about 784 Hz). At the highest f_o used (880 Hz), listeners could recognize the vowel /ɛ/ again.

The acoustic features and perceptual mechanisms underlying accurate vowel category perception at such high f_os remain unclear. As some of these studies found high identification rates even when excluding cues that play an important secondary role in vowel perception (e.g., vowel duration and formant frequency movement, see Lehiste and Peterson, 1961), it seems possible that spectral information apart from formant frequencies allowed listeners to identify vowels at very high f_os. Besides vowel identification models that are based on formant frequency distribution, speech scientists (in particular, from the automatic speech recognition community) have

long recognized that overall spectral shapes as reflected by, for example, Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980), are a more robust feature set than formants. Pols et al. (1969) and Klein et al. (1970) showed that a simple filter bank analysis (essentially an auditory excitation pattern approach which encodes the overall shape of the spectrum) matched perceptual vowel spaces well. Zahorian and Jagharghi (1993) found in an automatic vowel classification experiment that spectral-shape features (the discrete cosine transform coefficients of a bark frequency scaled spectrum) are superior acoustic cues for vowel identity classification compared to formants. Ito et al. (2001) showed that also the amplitude ratio of high- to low-frequency components (i.e., the spectral tilt) affects the perceived vowel category and is at least equally effective as F_2 as a cue for vowel identification. Several overall-spectral-shape models have been advocated over the last decades (see Kiefte et al., 2013, for a more comprehensive review of this approach). Most of them do not pay special attention to the distribution of formants, but are based on the assumption that the gross shape of a smoothed spectral envelope underlies the identification process. As it is very unlikely to find common formant frequency patterns at f_os of about 880 Hz, it seems possible that the overall spectral shape – despite a severe undersampling of the spectral envelope (see de Cheveigné and Kawahara, 1999, and Hillenbrand and Houde, 2003, for more details on this problem) – might have conveyed the information that allowed listeners to identify different vowel categories (but see Maurer, 2016, for an argument that perceived vowel categories are more a result of a complex systematic interaction between spectral shapes and f_o than has generally been assumed in phonetic theory).

However, it is also possible that the lack of between-talker acoustic vowel variation facilitated identification of the vowels (excepting Maurer and Landis, 1996, who used vowels of infant and adult talkers, all of the above-mentioned studies showing accurate vowel category perception at high f_os were single-talker studies). In that situation, listeners may have adapted to the talker's individual articulatory behavior (i.e., the within-talker acoustic vowel variation). Thus, it is not clear whether the results can be generalized to other talkers and whether an experimental design including more than one talker would lead to similar results. In addition, it seems likely that the number of response options (i.e., binary and multiple-choice tasks were used) had an effect on the identification performance as listeners perform better when fewer response options are provided.

The present study addresses these issues. Here, we asked three female talkers to produce the eight vowels /i y e ø ɛ a o u/ in isolation (thus eliminating possible confounding effects due to co-articulation with adjacent consonants) at eleven f_os within a range of 220–1046 Hz. In a multiple-choice task (mixed-talker condition) with all possible vowels as response options, listeners had to identify single 700-ms nuclei with quasi steady-state acoustic characteristics. These center portions of the vowels were used to exclude possible secondary cues, in particular, sweeping harmonics in the on- and off-sets, which might sample the vocal tract transfer function more continuously and thus provide information about the position of the formants.

To investigate possible spectral properties underlying listeners' identification process at high f_os, we calculated simple versions of the excitation patterns that these vowels would be expected to generate in the auditory periphery and discuss them with

respect to the results of the identification test.

4.3 Methods

4.3.1 Subjects

21 native German listeners (10 female, 11 male; mean age = 23.2, s.d. = 2.25) participated in a multiple-choice vowel identification task. All were students at the University of Zurich and none of them reported any hearing impairments when asked before the experiment.

4.3.2 Stimuli and apparatus

Three female native German talkers with professional voice training (one soprano, age: 33; one Musical-Theatre singer, age: 34; one actress, age: 34) were recorded with a cardioid condenser microphone (Sennheiser MKH 40 P48 with pop shield, Wedemark-Wennebostel, Germany) on a PC via an audio interface (RME Fireface UCX, RME, Halmhausen, Germany) in a noise-controlled room at Zurich University of the Arts (ZHdK) (Switzerland). The sampling frequency of the recordings was 44.1 kHz. Subjects were recorded keeping a constant distance of about 30 cm to the microphone when standing on a drawn position reference on the floor. They were selected based on samples from a corpus of recordings of 60 talkers because of their extended vocal range and noticeable skill of maintaining vowel categories at high f_os. As part of the standard procedure as implemented in an associated project (see Maurer et al., 2016, for more details), the latter was assessed in a listening test using a blocked-talker condition and a multiple-choice identification task carried out

by five phonetically trained listeners. The other 57 talkers (both female and male) had more limited vocal ranges and were not capable of producing vowels throughout the designated f_o range from 220 to 1046 Hz.

The three subjects were then asked to produce the eight long vowels /i y e ø ɛ a o u/ in isolation at eleven f_os (220, 330, 440, 523, 587, 659, 698, 784, 880, 988, 1046 Hz) with a monotone pitch contour resulting in 264 recordings (11 frequencies * 8 vowels * 3 talkers). Piano notes were presented as reference sounds to the subjects via loudspeaker immediately preceding the production. The talkers were asked to focus on producing recognizable vowels and to ignore typical voice aesthetics that might be important in their respective artistic style. The lowest f_o (220 Hz) corresponds to the female average f_o in citation-form words (Hillenbrand et al., 1995). The highest f_o (1046 Hz) corresponds to the high C (the musical note C6) in soprano singing and exceeds the normal range of F_1 of all German vowels produced by female talkers (see Pätzold and Simpson, 1997). The average f_o of each vowel was measured in Praat (Boersma and Weenink, 2016) using it's autocorrelation method (Boersma, 1993) and later checked manually. All vowels used in this study were recorded several times to ensure that at least one had an actual f_o close to the target f_o and a minimum duration of 1 second. All vowels that met these criteria were then evaluated again in the same listening test carried out by the five phonetically trained listeners, and the vowels with the highest identification scores were selected as stimuli. The mean duration of the final recordings was 1.49 s (range from on- to offset of voicing: 1.18–2.83 s).

Only vowel centers of 700 ms (± 350 ms from the vowel midpoint) with quasi-flat

f_o contours and steady-state spectral characteristics were used as stimuli. On- and offsets of the excised sounds were faded over 5 ms by amplitude modulating the waveform with raised cosines. All stimuli were normalized to an arbitrary intensity. The overall output level was chosen by listeners individually to be comfortable.

4.3.3 Procedure

A mixed-talker listening test was carried out in a small and noise-controlled room at the University of Zurich (Switzerland) using closed dynamic headphones (Beyerdynamic DT 770 Pro, 250 Ω). The experiment consisted of a multiple-choice identification task with all 8 vowels as response options. Listeners (n=21) were presented the excised 700-ms vowel nuclei while they saw a screen that contained eight circularly arranged buttons, each button labeled with one category (randomly arranged). Above the response buttons listeners could read the question *Welchen Vokal hörst Du?* (*Which vowel do you hear?*). The listener's task was to identify the vowel presented from the eight response options provided. After listeners made their choice they heard the next stimulus automatically with a delay of one second. Listeners could not repeat a stimulus. Each listener heard each token only once which means that any particular vowel at each f_o was responded to 63 times.

4.3.4 Data analysis

We performed a set of statistical analyses on correct/incorrect responses using mixed-effects logistic regression models in R (version 3.3.1; R Development Core Team, 2016, lmerTest package; Kuznetsova et al., 2014), in which listeners and items

were entered as random variables (Baayen et al., 2008). The predictors were vowel category, f_o, talker, and all their interaction. The significance of the main effects and interactions was assessed with likelihood ratio tests that compared the model with the main effect or interaction to a model without it. For clarity's sake, the results and figures are presented in percentages, although all statistical analyses were performed on raw data (correct/incorrect responses). The estimates (β) that are reported in the results section are expressed in logit units and were computed taking "incorrect response" as the reference level for the dependent variable.

To investigate possible shifts towards other than the intended vowel categories, 11 confusion matrices (one for each f_o, each based on a total of 504 samples, i.e., 8 vowels x 3 talkers x 21 listeners' responses) with the two dimensions *intended vowel* (actual class) and *response vowel* (predicted class) were calculated.

4.3.5 Excitation patterns

Simple auditory excitation patterns were generated for each vowel using a 200-channel linear gammatone filter bank, whose bandwidths and centre frequencies were calculated according to the ERB formulae given by Glasberg and Moore (1990). The rms level of the output wave was calculated for each filter channel, and converted to dB. In addition, a frequency weighting was applied to account for the transmission properties of the middle ear, as based on measurements made by Puria et al. (1997).

4.4 Results

Results obtained from the logistic regression revealed a highly significant effect of f_o ($\chi^2(10) = 30.8$, $p < .001$), a highly significant effect of vowel category ($\chi^2(7) = 28.21$, $p < .001$), no main effect of talker ($\chi^2(2) = 2.24$, $p = .33$), and a highly significant interaction between the three ($\chi^2(244) = 627.91$, $p < .001$). For the ease of interpretation, and as a complex three-way interaction makes it impossible to ignore any one of them in accounting for the effects of the other two, we decided to break down the data into three sets to test for a two-way interaction between vowel category and f_o for the individual talkers. The results of the three analyses showed consistently a highly significant interaction between vowel category and f_o (talker 1: $\chi^2(70) = 188.42$, $p < .001$; talker 2: $\chi^2(70) = 182.74$, $p < .001$; talker 3: $\chi^2(70) = 209.5$, $p < .001$). Significant effects of vowel category were found for all talkers (talker 1: $\chi^2(7) = 28.19$, $p < .001$; talker 2: $\chi^2(7) = 22.01$, $p < .01$; talker 3: $\chi^2(7) = 35.77$, $p < .001$), and f_o (talker 1: $\chi^2(10) = 30.79$, $p < .001$; talker 2: $\chi^2(10) = 32.61$, $p < .001$; talker 3: $\chi^2(10) = 30.2$, $p < .001$). Taken together, these effects suggest that listeners' identification performance showed high variability between vowel categories and across f_os generally.

Figure 4–1 shows the distribution of the percentage of correct identification for each f_o and talker across vowels. Throughout the f_o range the overall performance declined more or less continuously for all talkers.

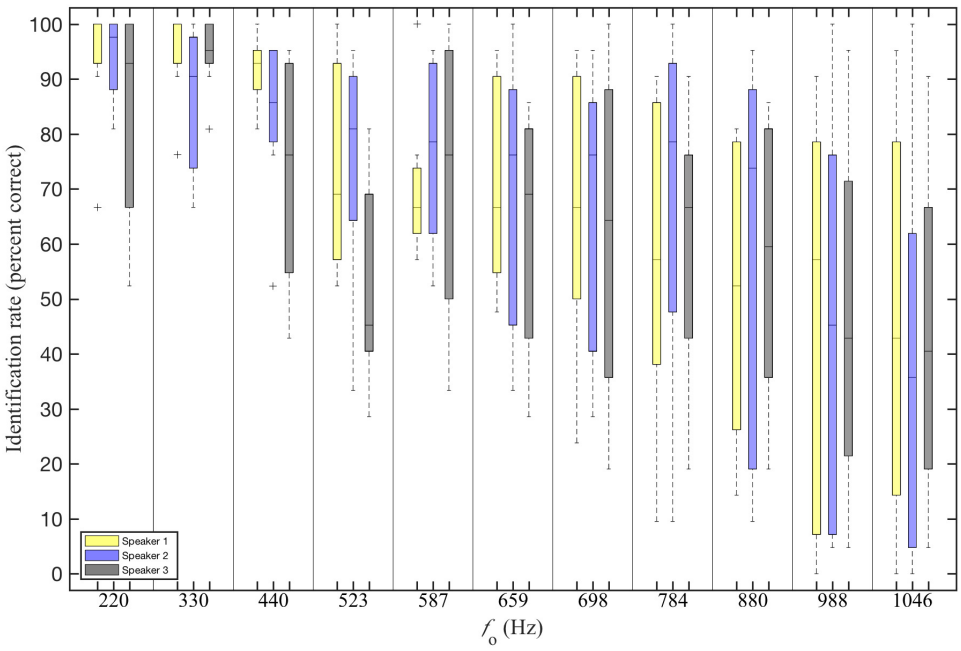

Figure 4–1: (Color online) Box plots showing the distribution of percent correct for the identification of all investigated vowels at the eleven f_os for the individual talkers.

The increasing variability toward the higher f_os can be explained by an increasing inter-vowel variability, as the identification rate of individual vowel categories differed greatly between low and high f_os. This can be seen in Figure 4–2 showing the mean percent correct scores for each individual vowel at the different f_os. Listeners' identification performance for the vowels /i ɛ a u/ is surprisingly stable up to at least 880 Hz, and percent correct values can typically be found in the range above 70%. At the two highest f_os (988 and 1046 Hz), the identification rate for /ɛ/ drops to intermediate ranges between 40 and 50% correct. Only the point vowels /i a u/ remain in the upper third of the percent correct scale. On the contrary, for

the vowels /e ø o/ an extensive decrease in listeners' identification performance can be found throughout the f_os from 220 to 1046 Hz. While identification scores range between 90–100% at the two lowest f_os (220 and 330 Hz), they drop fairly continuously toward chance level for these three vowels, which is reached at 988 Hz. The identification rate of /y/ drops substantially at an f_o of 523 Hz (from about 85 to 60% correct) and decreases despite some variability towards upper f_os. From 988 Hz identification scores are similar to those of /ɛ/ (i.e., within the 35–50% correct range).

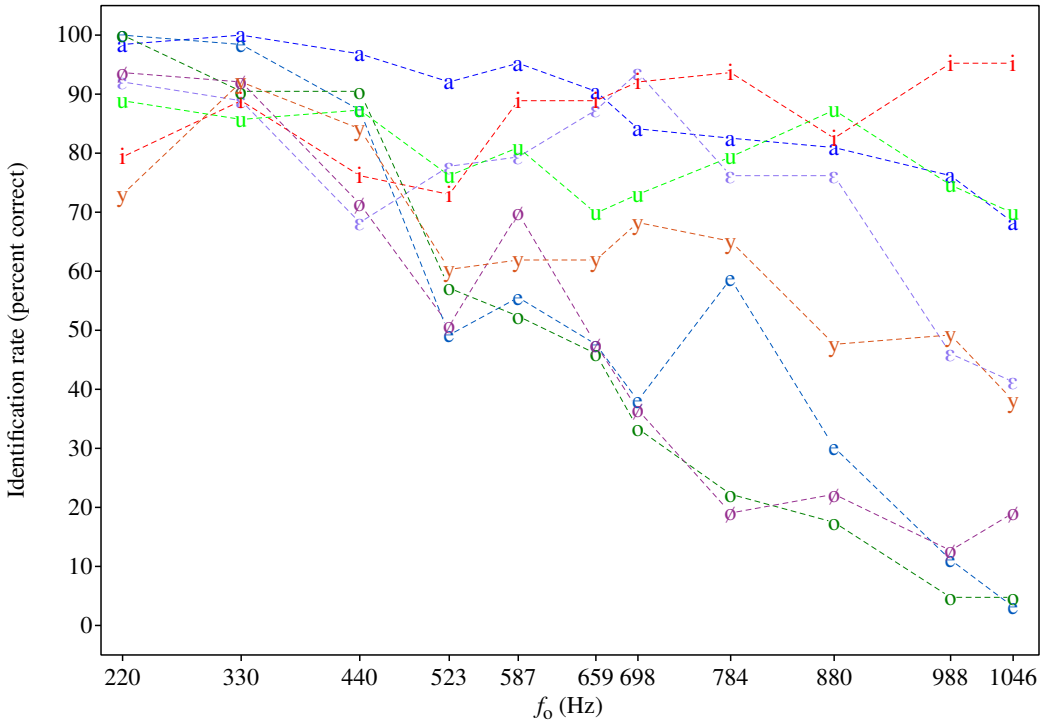

Figure 4–2: (Color online) Line graphs showing percent correct values, summed over all talkers, for the identification of each of the eight vowels over the investigated f_o range.

Confusion matrices (see Fig. 4–3, for a graphical illustration; the raw data can be found in Appendix A) reveal dominant shifts toward the vowel categories /i a u/ in cases of false identifications at the highest f_os. For /ɛ/, strong confusions at the highest two f_os (988 and 1046 Hz) were found with /a/, which also showed the highest response proportions of all vowels at these f_os (28% and 24.4%). The drop in identification performance for the vowel /y/ in the range from 523 Hz on upwards is due to a confusion with other front vowels and from 784 Hz upwards mainly due to a confusion with /i/. A confusion between these two vowels also explains the relatively poor performance for /i/ at the lowest f_o 220 Hz (15.9% of the listeners responded /i/ when /y/ was presented to them). In case of /ø/, shifts in perception were generally found to be widely spread, that is, toward all the investigated vowel categories except /i/. The majority of false identification of /o/ shifted from a perceived /a/ at 523 and 587 Hz to /u/ at all higher f_os. Within the range 523–784 Hz, the vowel /e/ was often confused with /i/. At higher f_os the perceived vowel category shifted toward /ɛ/ and /a/.

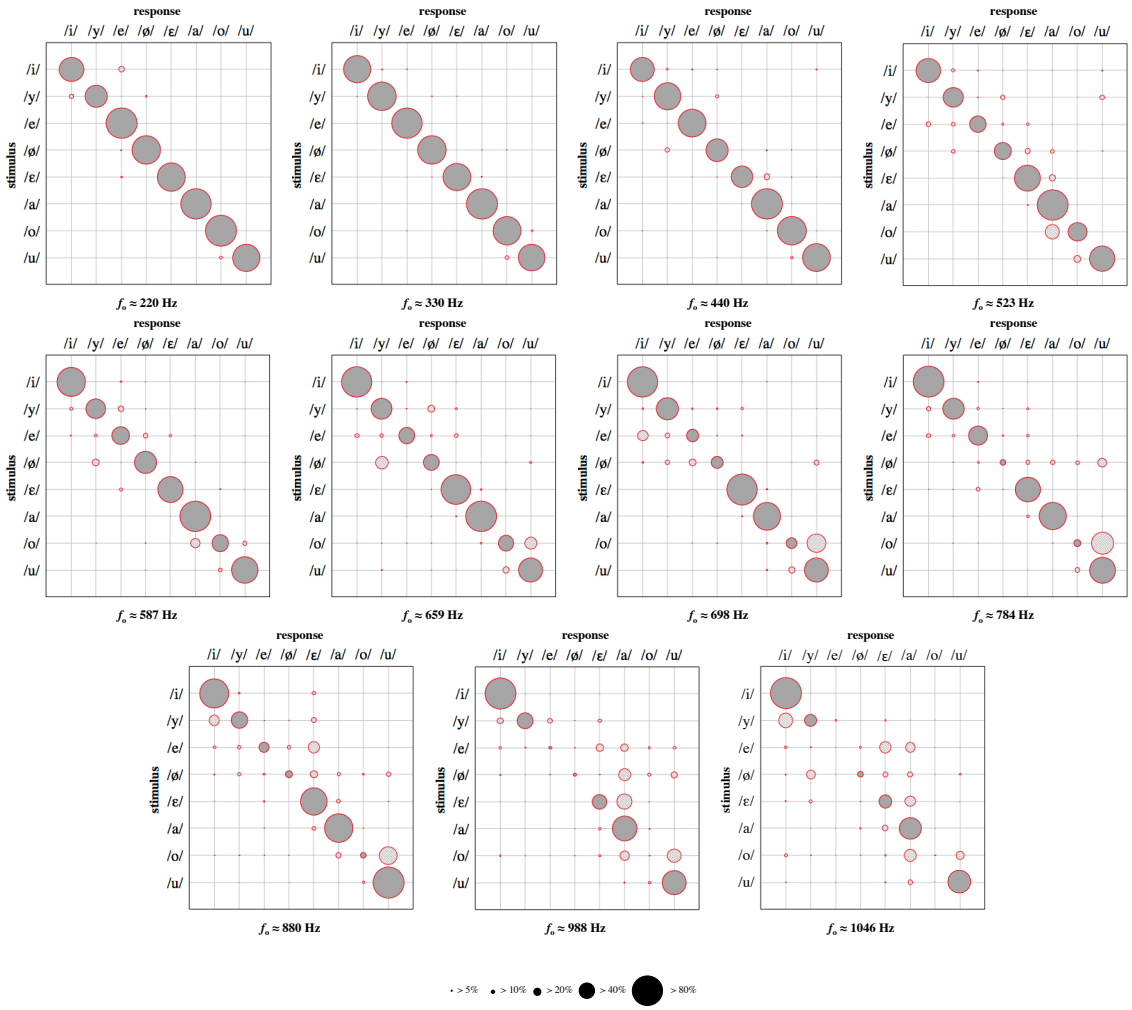

Figure 4–3: (Color online) Graphical confusion matrices showing the intended and response vowel categories for each f_o. The radius of each circle is proportional to the number of times that a particular stimulus (given by the row) was identified as the column response. Correct responses (down the diagonal) are solid gray, whereas identification errors (confusions) are indicated by diagonal lines through the circles.

Figure 4–4 shows the auditory excitation patterns for the eight vowels used in this study produced at an f_o of about 988 Hz. Both the patterns calculated for

individual talkers and those averaged across talkers reveal that the point vowels /i a u/ show maximally distinct spectral shapes, which can be easily distinguished by the overall excitation level in the higher frequency region above about 1.5 kHz. The obtained confusions of the vowel categories /y e ø ɛ o/ at this f_o show a high degree of correspondence to the excitation patterns of the respective point vowels they were confused with most often. For example, the pattern calculated for /o/ shows high similarity with the pattern of the point vowel /u/, that is, a relatively low excitation level in the high-frequency region. The excitation pattern of /y/ exhibits a relatively high excitation level in the high-frequency region, which is also the case for the point vowel /i/. The patterns of the vowels /e ø ɛ/ show intermediate levels of excitation in the high-frequency region, which is also the case for /a/, the vowel which was most often responded by the listeners when these vowels were presented to them at 988 Hz.

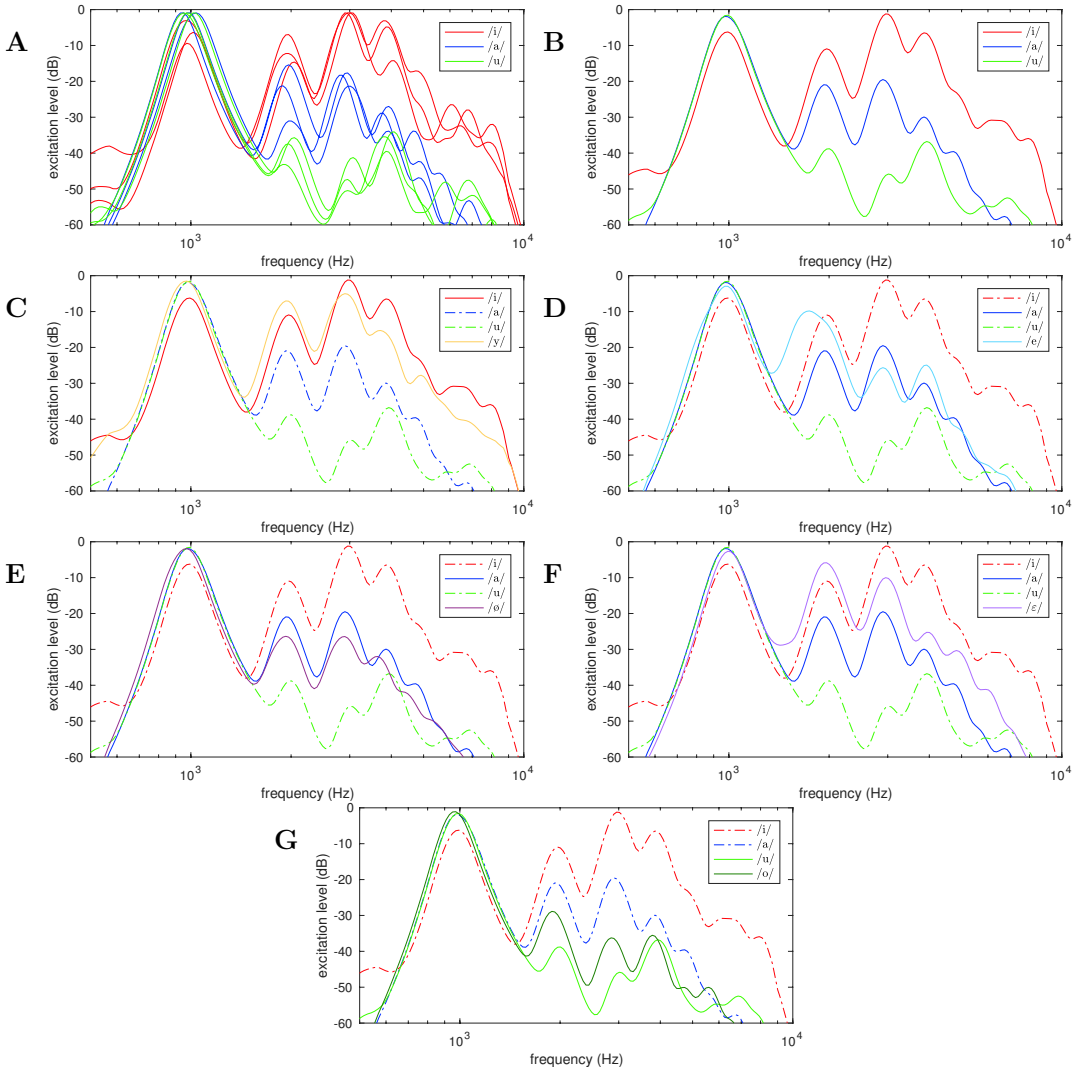

Figure 4–4: (Color online) Excitation patterns for the vowels used in this study that had an f_o of about 988 Hz. Part (A) shows the excitation patterns for the individual point vowels /i a u/ produced by all talkers. Part (B) shows the excitation patterns of the same vowels averaged across talkers. All other parts (C–G) show each of the other investigated vowels together with the point vowels. In these graphs, solid lines are used to indicate the strongest confusion of a respective vowel with one of the point vowels. (The information in this figure may not be properly conveyed in black and white.)

4.5 Discussion

The results have shown that listeners' abilities to recognize vowels within a fundamental frequency range from 220 to 1046 Hz differ greatly across vowel categories and the range of f_os. Listeners could perform well even with a variety of talkers, which means that good performance at high f_os is not being done through some odd mechanism or sensitivity which would be idiosyncratic for each talker. It is not surprising that all vowels could be identified accurately at the lowest f_os used here (220 and 330 Hz), but it is striking that only the performance for the vowels /y e ø o/, but not for /i a ɛ u/ decreased drastically within the f_o range from around 523 to 880 Hz. The results also revealed that the point vowels /i a u/ remain identifiable at an f_o close to 1 kHz or even above (in the case of /i/).

Thus, the results differ substantially from those provided by numerous studies on vowel identification in Western classical singing, which have reported consistently that high vowels such as /i/ and /u/ are the first vowels to lose their identity when f_o is progressively increased. This means that findings from the field of operatic singing cannot be generalized to other forms of speech production. In addition, the findings reported here support the hypothesis that articulatory changes which have been found in Western classical singers like resonance tuning (e.g., shifting f_{R1} to the vicinity of a higher f_o), must indeed have a strong effect on the identifiability of vowels.

Given the degree to which the vocal tract transfer function is undersampled at an f_o around 1 kHz a significant loss of formant information has to be considered as

very likely (e.g., here, the vowels' typical medians of F_1 are exceeded by about 220–660 Hz, and there is only one harmonic every 1 kHz). Although it is possible that the loss of formant information can explain the decreasing identification performance, it seems likely that formants cannot be the primary acoustic correlates for vowel category perception at very high f_os.

Calculations of auditory excitation patterns for the eight vowels at an f_o of 988 Hz, revealed maximally distinct excitation levels in the frequency region above roughly 1.5 kHz for the point vowels /i a u/. Excitation patterns of the other vowels have been found to exhibit very similar spectral shapes as those of the point vowels they have been confused with most often. Both the excitation patterns of /u/ and /o/, for example, show relatively low excitation in the frequency region above 1.5 kHz, but the identification rate of /u/ (about 75% correct) was considerably higher than that of /o/ (about 10% correct), while a substantial proportion of responses (about 43%) were /u/ when /o/ was presented. As similar observations were found for other non-point and point vowel combinations, it seems likely that distinctive excitation patterns can be used by listeners as landmarks (in terms of reference points) for vowel category perception at high f_os.

Using distinctive excitation patterns as landmarks for vowel identification could also explain most of the findings reported in earlier studies on vowel identification at high f_os. Regarding the vowels used by Smith and Scott (1980) in their perception experiment (i.e., /i ɪ ɛ æ/), it is possible that the information conveyed by the distinct spectral shapes might have been sufficient for the listeners to distinguish at least between the two pairs /i ɪ/ and /ɛ æ/. However, it is difficult to draw conclusions

from this as vowel duration differed substantially in this study, and not enough detail about performance with the different vowels and the instructions given to the listeners were provided.

Comparing the results of the present study to those reported by Friedrichs et al. (2015b), the diverging identification performance for the vowel /o/ is surprising. While a perfect identification rate (100% correct) was found at an f_o of 880 Hz by Friedrichs et al. (2015b), a performance near chance (17.5% correct) was observed in the present study. Although the lack of between-talker acoustic vowel variation (as being a single talker study) and secondary cues to vowel identity (vowels were presented in word context) in the former study might have helped listeners to perform better it seems possible that this difference is also due to the importance of perceptual and acoustic landmarks. The strongest support for this hypothesis is the fact that the vowel /u/ was not included in the study of Friedrichs et al. (2015b), and thus, a confusion of /o/ and /u/ like the one found in the present study was not possible (e.g., /u/ received more than 50% of the responses for the intended vowel /o/ at an f_o of 880 Hz). It seems, therefore, likely that listeners used the vowel /o/ as a substitute because /u/ was not presented to them as a response option. The results by Friedrichs et al. (2015a), who found the same eight vowels used in the present study identifiable up to an f_o of 880 Hz when recorded in minimal pairs and tested in a two-alternative forced choice task, could also be explained within this context. As a single talker was asked to produce several different two-word combinations containing a vowel in contrastive position (e.g., the German words *Buden* vs. *Boden*), it is possible that the talker produced vowels with acoustic features alike or different from

those of a point vowel at higher f_os to make them distinguishable (e.g., producing an /o/ more toward /a/ to distinguish it from /u/). This way the phonological function of vowels in linguistic contrastive positions could be maintained for all vowels even at very high f_os. Given this, it is plausible that the number of response options has a strong effect on listeners' identification performance, and obviously, a better performance should be expected when fewer responses options are provided.

It is possible that the results presented here may have been driven in part by the relative frequency of German vowels. For example, in German, /i/ is more frequent than /y/, and /u/ is more frequent than /o/ (Pätzold and Simpson, 1997). Forced to choose between two vowels that otherwise match the spectral characteristics of the stimulus equally well, listeners are most likely to pick the one with the higher a priori probability. However, it is unlikely that this can explain listeners' identification performance entirely as, for example, the long /e/ is more frequent than the long /a/, with which it has been confused most often in this study at an f_o of 988 Hz. In addition, relative frequency may be the driving force behind which vowel label is applied to a cluster of similar vowels, but it cannot explain the fact that vowels were categorized into three distinct groups.

In summary, the results presented here make it clear that a theory of vowel perception based solely on formant peak patterns cannot account for the relatively preserved performance listeners demonstrate in identifying vowels at high f_os. Formal modelling of the relationship between the perceptual and physical spaces of vowels at high and low f_os are required for a convincing demonstration, but it seems likely that overall spectral shape features will play an important role in a coherent account

of vowel perception generally.

4.6 Acknowledgements

This study was supported by the Forschungskredit of the University of Zurich, Grant No. FK-14-062, and the Swiss National Science Foundation (SNSF), Grants No. P2ZHP1_168375 and 100016_143943/1. Thanks to Nick Clark, whose software was used to perform the gammatone filtering, and Sandra Schwab for her helpful contributions and comments on an earlier draft of this paper.

4.7 References

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**(4), 390–412.

Boersma, P., and Weenink, D. (**2016**). "Praat: Doing phonetics by computer [computer program]," Version 6.0.15, retrieved March 23, 2016 from http://www.praat.org/ (Last viewed April 30, 2016).

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. Inst. Phonetic Sci. (17), University of Amsterdam, 97–110.

Davis, S., and Mermelstein, P. (**1980**). Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. Proc. IEEE Int. Conf. Acoust. Speech Signal Processing **28**, 357–366.

de Cheveigné, and Kawahara, H. (**1999**). "Missing-data model of vowel identification," J. Acoust. Soc. Am. **105**, 3497–3508.

Friedrichs, D., Maurer, D., and Dellwo, V. (**2015a**). "The phonological function of vowels is maintained at fundamental frequencies up to 880Hz," J. Acoust. Soc. Am. **138** , EL36–EL42.

Friedrichs, D., Maurer, D., Suter, H., and Dellwo, V. (**2015b**). "Vowel identification at high fundamental frequencies in minimal pairs," Proceedings of the 18th Internaional Congress of Phonetic Sciences, paper number 0438, 1–5.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," J. Acoust. Soc. Am. **113**, 1044–1055.

Howie, J., and Delattre, P. (**1962**). "An experimental study ofthe effect of pitch on the intelligibility of vowels," Natl. Assoc. Teachers Singing Bull. **18**(4), 6–9.

Ito, M., Tsuchida, J., and Yano, M. (**2001**). "On the effectiveness of whole spectral shape for vowel perception," J. Acoust. Soc. Am. **110**(2), 1141–1149.

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). "Vocal tract resonances in singing: the soprano voice," 116, 2434–2439.

Kiefte, M., Neary, T. M., and Assmann, P. F. (**2013**). "Vowel Perception in Normal Speakers," Handbook of Vowels and Vowel Disorderes, edited by M. J. Ball, and F. E. Gibbon (Taylor and Francis, New York), pp. 161–185.

Klein, W., Plomp, R., and Pols, L. C. (**1970**). "Vowel spectra, vowel spaces, and vowel identification," J. Acoust. Soc. Am. **48**(4B), 999–1009.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2014**). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0–20. Retrieved from http://CRAN.R-project.org/package=lmerTest. (Accessed on June 30, 2016)

Lehiste, I., and Peterson, G. E. (**1961**). "Transitions, glides, and diphtongs," J. Acoust. Soc. Am. **33**, 268–277.

Maurer, D., and Landis, T. (**1996**). "Intelligibility and spectral differences in high-pitched vowels," Folia Phoniatr. Logop. **48**, 1–10.

Maurer, D., Mok, P., Friedrichs, D., and Dellwo, V. (**2014**). "Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer," Proceedings of the Fifteenth Annu. Conf. Int. Speech Commun. Assoc. Singapore, 2132–2133.

Maurer, D., Suter, H., Friedrichs, D., and Dellwo, V. (**2016**). "Acoustic characteristics of voice in music and straight theatre: topics, conceptions, questions," Trends in Phonetics and Phonology. Studies from German speaking Europe, edited by A. Leemann, M. J. Kolly, S. Schmid, and V. Dellwo (Peter Lang, Bern/Frankfurt), pp. 256–265.

Maurer, D. (**2016**). *Acoustics of the Vowel - Preliminaries* (Peter Lang AG, International Academic Publishers, Bern).

Pätzold, M., and Simpson, A. (**1997**). "Acoustic analysis of German vowels in the Kiel Corpus of read speech," Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel **32**, 215–247.

Pols, L. C., Van der Kamp, L. T., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**(2B), 458–467.

Puria, S., Peake, W. T., and Rosowski, J. J. (**1997**) Sound-pressure measurements in the cochlear vestibule of human cadaver ears, J. Acoust. Soc. Am. **101**, 2754–2770.

R Core Team. (**2016**). R: A Language and Environment for Statistical Computing. Version 3.1.3. [Computer software] Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org (Accessed on June 30, 2016)

Smith, J., and Wolfe, J. (**2009**). "Vowel-pitch matching in Wagner's operas: Implications for intelligibility and ease of singing," J. Acoust. Soc. Am. **125**, EL196–EL201.

Smith, L. A., and Scott, B. L. (**1980**). "Increasing the intelligibility of sung vowels," J. Acoust. Soc. Am. **67**, 1795–1797.

Sundberg, J. (**1975**). "Formant technique in a professional female singer," Acustica **32**, 89–96.

Sundberg, J. (**2013**). "Perception of singing," in Psychology of Music, 3rd ed., edited by D. Deutsch (Academic Press, London), pp. 69–106.

Zahorian, S., and Jagharghi, A. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–82.

4.8 Appendix

See Table 4–1

Table 4–1: Confusion matrices for each f_o containing the raw data of the identification test in percentages.

$f_o \approx 220$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	79.4	0	20.6	0	0	0	0	0
/y/	15.9	73	3.2	7.90	0	0	0	0
/e/	0	0	100	0	0	0	0	0
/ø/	0	0	6.3	93.7	0	0	0	0
/ɛ/	0	0	7.9	0	92.1	0	0	0
/a/	0	0	0	0	1.6	98.4	0	0
/o/	0	0	0	0	0	0	100	0
/u/	0	0	0	0	0	0	11.1	88.9
response proportions	11.9	9.10	17.3	12.7	11.7	12.3	13.9	11.1

$f_o \approx 330$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	6.3	4.8	0	0	0	0	0
/y/	4.8	92.1	0	1.6	1.6	0	0	0
/e/	1.6	0	98.4	0	0	0	0	0
/ø/	0	0	0	92.1	0	4.8	3.2	0
/ɛ/	0	0	3.2	1.6	88.9	6.3	0	0
/a/	0	0	0	0	0	100	0	0
/o/	0	0	1.6	0	0	0	90.5	7.9
/u/	0	0	0	0	0	0	14.3	85.7
response proportions	11.9	12.3	13.5	11.9	11.3	13.9	13.5	11.7

$f_o \approx 440$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	76.2	7.9	6.3	4.8	0	0	0	4.8
/y/	4.8	84.1	0	11.1	0	0	0	0
/e/	4.8	1.6	87.3	3.2	3.2	0	0	0
/ø/	0	15.9	0	71.4	3.2	6.3	3.2	0
/ɛ/	0	0	1.6	4.8	68.3	20.6	3.2	1.6
/a/	0	0	0	0	1.6	96.8	1.6	0
/o/	1.6	0	0	0	0	4.8	90.5	3.2
/u/	0	1.6	0	1.6	0	0	9.5	87.3
response proportions	10.9	13.9	11.9	12.1	9.5	16.1	13.5	12.1

$f_o \approx 523$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	73	11.1	6.3	1.6	0	1.6	0	6.3
/y/	1.6	60.3	4.8	15.9	0	0	1.6	15.9
/e/	15.9	12.7	49.2	7.9	9.5	3.2	0	1.6
/ø/	0	12.7	1.6	50.8	17.5	12.7	1.6	3.2
/ɛ/	0	0	0	1.6	77.8	20.6	0	0
/a/	0	0	0	0	4.8	92.1	3.2	0
/o/	0	0	0	0	0	42.9	57.1	0
/u/	0	0	0	0	0	1.6	22.2	76.2
response proportions	11.3	12.1	7.7	9.7	13.7	21.8	10.7	12.9

$f_o \approx 587$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	0	7.9	1.6	0	0	0	1.6
/y/	12.7	61.9	19	4.8	0	1.6	0	0
/e/	6.3	11.1	55.6	15.9	7.9	1.6	0	1.6
/ø/	0	22.2	1.6	69.8	0	4.8	0	1.6
/ɛ/	0	0	11.1	0	79.4	0	6.3	3.2
/a/	0	0	0	0	1.6	95.2	3.2	0
/o/	0	1.6	0	1.6	0	30.2	52.4	14.3
/u/	0	0	0	1.6	0	3.2	14.3	81
response proportions	13.5	12.1	11.9	11.9	11.1	17.1	9.5	12.9

$f_o \approx 659$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	88.9	1.6	4.8	0	3.2	0	0	1.6
/y/	3.2	61.9	4.8	20.6	7.9	0	0	1.6
/e/	14.3	11.1	47.6	7.9	14.3	0	3.2	1.6
/ø/	0	38.1	1.6	47.6	1.6	1.6	1.6	7.9
/ɛ/	0	0	0	3.2	87.3	7.9	1.6	0
/a/	0	0	1.6	1.6	6.3	90.5	0	0
/o/	1.6	3.2	3.2	3.2	0	6.3	46	36.5
/u/	0	4.8	0	1.6	1.6	1.6	20.6	69.8
response proportions	13.5	15.1	8	10.7	15.3	13.5	9.1	14.9

$f_o \approx 698$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	92.1	0	3.2	0	1.6	3.2	0	0
/y/	6.3	68.3	6.3	7.9	9.5	0	0	1.6
/e/	33.3	15.9	38.1	4.8	6.3	0	0	1.6
/ø/	7.9	14.3	22.2	36.5	0	0	1.6	17.5
/ɛ/	0	0	0	0	93.7	6.3	0	0
/a/	0	1.6	3.2	3.2	6.3	84.1	1.6	0
/o/	0	0	1.6	1.6	0	6.3	33.3	57.1
/u/	0	0	0	0	0	6.3	20.6	73
response proportions	17.5	12.5	9.3	6.8	14.7	13.3	7.1	18.9

$f_o \approx 784$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	93.7	0	4.8	0	1.6	0	0	0
/y/	15.9	65.1	9.5	1.6	7.9	0	0	0
/e/	14.3	9.5	58.7	6.3	9.5	0	1.6	0
/ø/	0	3.2	7.9	19	14.3	14.3	12.7	28.6
/ɛ/	4.8	3.2	12.7	3.2	76.2	0	0	0
/a/	0	1.6	1.6	0	9.5	82.5	3.2	1.6
/o/	0	3.2	1.6	0	0	4.8	22.2	68.3
/u/	0	0	0	0	1.6	3.2	15.9	79.4
response proportions	16.1	10.7	12.1	3.8	15.1	13.1	7	22.2

$f_o \approx 880$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	82.5	6.3	0	0	11.1	0	0	0
/y/	30.2	47.6	3.2	3.2	15.9	0	0	0
/e/	9.5	11.1	30.2	11.1	33.3	3.2	0	1.6
/ø/	4.8	11.1	7.9	22.2	22.2	11.1	6.3	14.3
/ɛ/	1.6	0	6.3	0	76.2	12.7	0	3.2
/a/	0	0	3.2	0	11.1	81	3.2	1.6
/o/	3.2	4.8	3.2	4.8	0	15.9	17.5	50.8
/u/	0	1.6	0	1.6	0	1.6	7.9	87.3
response proportions	16.5	10.3	6.8	5.4	21.2	15.7	4.4	19.9

$f_o \approx 988$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	95.2	1.6	1.6	0	1.6	0	0	0
/y/	20.6	49.2	15.9	1.6	12.7	0	0	0
/e/	9.5	6.3	11.1	4.8	23.8	25.4	7.9	11.1
/ø/	6.3	1.6	4.8	12.7	4.8	38.1	11.1	20.6
/ɛ/	1.6	1.6	0	0	46	47.6	3.2	0
/a/	0	0	3.2	1.6	9.5	76.2	6.3	3.2
/o/	6.3	1.6	3.2	3.2	7.9	30.2	4.8	42.9
/u/	3.2	3.2	1.6	0	1.6	6.3	9.5	74.6
response proportions	17.8	8.1	5.2	3	13.5	28	5.4	19.1

$f_o \approx 1046$ Hz	/i/	/y/	/e/	/ø/	/ɛ/	/a/	/o/	/u/
/i/	95.2	1.6	0	0	3.2	0	0	0
/y/	44.4	38.1	7.9	0	6.3	1.6	1.6	0
/e/	9.5	6.3	3.2	7.9	36.5	31.7	3.2	1.6
/ø/	6.3	28.6	1.6	19	17.5	17.5	1.6	7.9
/ɛ/	6.3	11.1	0	4.8	41.3	33.3	0	3.2
/a/	0	3.2	1.6	6.3	19	68.3	1.6	0
/o/	11.1	4.8	3.2	4.8	6.3	38.1	4.8	27
/u/	4.8	1.6	1.6	0	4.8	15.9	1.6	69.8
response proportions	22.2	11.9	2.4	5.4	16.9	25.8	1.8	13.7

CHAPTER 5

Study IV

Methodological issues in the acoustic analysis of steady state vowels

The following article appeared in

Trends in Phonetics and Phonology. Studies from German speaking Europe, 33–41,

and may be found at

D. Friedrichs, D. Maurer, H. Suter, and V. Dellwo, "Methodological issues in the acoustic analysis of steady state vowels," in: A. Leemann, M.-J. Kolly, S. Schmid, and V. Dellwo (Ed.) *Trends in Phonetics and Phonology. Studies from German speaking Europe*, 33–41 (**2015**).

5.1 Abstract

Formant frequency analysis represents the current standard for determining speech-specific acoustic characteristics of vowel sounds: For souneds with quasi-constant spectral characteristics, according to source-filter theory, vowel quality relates to vowel-specific formant patterns. With regard to the determination of formant frequencies, -bandwidths, and -amplitudes, current methods, in general, rely on linear prediction (LP) and on visual inspection of spectrograms. Both methods require a high degree of experience and expertise. Above all, first, inappropriate selection of parameter settings for LP analysis often produces unreliable numerical values in general; second, cross-check of calculated formant values on the basis of spectrograms is limited, thus makes formant analysis of large samples of vowel sounds difficult; and third, severe difficulties in formant measurement occur in cases of sounds exhibiting high fundamental frequencies. The present paper discusses the basic aspects of the two analysis methods mentioned, and relates these aspects to the investigation of large databases and the investigation of an extensive variation of fundamental frequency.

5.2 Introduction

With regard to the analysis of the speech-related acoustic characteristics of vowel sounds, the measurement of two properties are probably the most salient in terms of their perception: Fundamental frequency (f_o), which is the acoustic correlate of pitch, and formant frequencies (in particular F_1, F_2, F_3), the acoustic correlates of vowel quality. The importance of these two elements becomes clear in a classical

concept of human speech production, which assumes that vowel sounds are determined by the acoustic characteristics of the source signal and the resonances of the vocal tract configuration (Chiba and Kajiyama, 1941). From this approach the well-known source-filter theory evolved, which is mostly associated with the works of Stevens and House (1955), and Fant (1960). According to the source-filter theory, concerning voiced sounds, an air stream from the lungs is modulated into quasi-periodic air pulses by the vocal folds, which repeatedly open and close. Through this glottal opening/closing action, a characteristic airflow signal is produced at the lower end of the vocal tract. This creates a complex acoustic wave that determines the fundamental frequency (f_o; the number of periods, i.e. pulses, per second) and its integer multiples as a harmonic spectrum. For normal voice production, the amplitude of the harmonics rolls off by about 6–18 dB/octave (Mathews, 1999; Baken and Orlikoff, 2000). This spectral roll-off varies with vocal effort. When speaking loud (high effort), the harmonics roll off less strongly, in contrast, when speaking with a breathy voice (low effort), the harmonics roll off stronger. The source signal is then filtered by the resonances of the vocal tract in relation to varying positions of the velum, tongue, jaw, and lips (for singing and acting, in addition, lowering the larynx and/or narrowing the epipharyngeal tube have to be considered; Imagawa et al., 2003; Bele, 2006). This means that the harmonics of the source spectrum are either attenuated by the resonances of the vocal tract or may pass it relatively unchanged in amplitude. This process determines a characteristic spectral shape that varies between different vowels of a language. According to the definition by Fant (1960,

p. 20), the formants are "the spectral peaks of the sound spectrum", each peak representing a particular resonance of the vocal tract. Results from numerous empirical studies indicate that vowel quality relates to vowel-specific formant patterns if the corresponding sounds are investigated at f_o of 'normal speech' (see, e.g., Petersen and Barney, 1952; Fant, 1959; Hillenbrand et al., 1995). Furthermore, results from speech synthesis research support these findings. Most vowels can be synthesized applying a two-formant synthesis, and a three-formant synthesis allows for a perceptual distinction of all vowels in a given language (Bladon and Fant, 1978; Kent and Read, 2002). However, such findings are limited for vowel syntheses not including an extensive variation of f_o (see section 5.5). Although commonly used, current methods to track f_o (e.g., time domain detection, frequency domain detection) and to determine formant patterns (linear prediction and spectrographic depiction) are still error-prone. Choosing parameter settings for the acoustic analysis is in many cases not an easy task as the definition of thresholds might be appropriate for some but not for other vowels. In addition, utterances including an extensive variation of the fundamental frequency (e.g., f_o exceeding F_1 values given in the literature) may cause problems. Therefore, the present paper describes and illustrates major methodological issues in the acoustic analysis of vowel sounds. To clarify basic problems, we focus on acoustic analysis of steady state vowel sounds (with quasi-constant spectral characteristics and exclusion of transitions) carried out in Praat (Boersma and Weenink, 2014).

5.3 Fundamental frequency measurement

With all commonly used analysis tools, f_o tracking is typically conducted via an autocorrelation method. This method correlates a part of the signal with a so-called lag part, i.e. a part occurring after a certain time lag. The analysis provides an estimate of the periodicity (number of periods per second) of the signal. However, the method may produce typical artifacts. Above all, a halving or doubling of the measurement values may occur: For example, f_o analysis of a sound with an actual periodicity of 300 Hz may result in either 150 or 600 Hz. This phenomenon is also known as 'octave jumps' or 'octave errors'. The Praat autocorrelation algorithm (based onBoersma, 1993) calculates almost consistently accurate f_o values during steady state vowels. However, also in this environment octave errors can occur. In such cases, false tracking can only be intercepted by visual inspection of the spectrum (e.g., determining the frequency of the first harmonic) or by measuring the duration of a period of the sound wave. Yet, this is only practicable when investigating a rather small number of sounds. When building up a large database of steady state vowels, e.g., for the analysis of vocalic variability in singing and speech, these approaches are relatively laborious and time-consuming. We, therefore, developed a practical way to immediately discover octave errors already during the recording. In this, f_o is provided (played back acoustically) to the speakers in form of a reference sound. In addition, the investigator edits f_o ranges immediately when the recordings are carried out. For each utterance recorded, thus, reference values or narrow frequency ranges of such references are determined. If these reference values differ substantially from the calculated values, corrections can be made for each single

sound on the basis of inspection of the harmonics in the spectrum. The same can be applied when, in addition to a large sample of sounds, extensive variation of f_o is also investigated. f_o analysis becomes more difficult when the source sound exhibits substantial non-periodic acoustic characteristics, as is the case, e.g., for creaky or breathy voices. For the related sounds, analysis relies on the actual characteristics of phonation and must be considered correspondingly (see, e.g., Hillenbrand et al., 1994).

5.4 Formant analysis

In general, recent studies make use of two methods for estimating formant patterns of vowel sounds: Linear prediction analysis (LP analysis, or synonymously, linear predictive coding [LPC]), and spectrographic depiction. Moreover, many studies link these two methods together, that is, calculation of numerical values of frequencies, bandwidths, and amplitudes of the formants is carried out by LP analysis, and these values are crosschecked by visual inspection of the related spectrogram. LP analysis relies on the source-filter theory of speech production. Simply put, it is based on a decomposition of a sound wave into a source and a filter, where the filter shape is assumed to correspond to the vocal tract resonances. As a result, values for each formant can be derived from a calculated filter curve that represents the transfer function of the vocal tract. For spectrographic depiction, a Fourier Transform (e.g., fast Fourier Transform [FFT]) needs to be performed. A good way to estimate formant frequencies is to use a wide-band spectrogram, showing frequency vs. time, with intensity as darkness. Thus, in the spectrogram, frequency ranges of the

highest energy (darkest bars) correspond to formants. Both the LP analysis and the spectrographic estimation have advantages and disadvantages in terms of formant pattern estimation, which are discussed here on the basis of a practical approach in Praat.

5.4.1 Linear prediction in Praat

Praat allows the possibility of choosing between different algorithms that are all based on linear prediction (LP). This includes algorithms that are integrated into the commands 'To LPC...' and 'To Formant...' (and additional sub-commands). In general, LP requires different parameters/coefficients that are either given to the particular algorithm or have to be chosen by the investigator: (1) Time step(s) to determine the frames for which analysis will be carried out within the total duration of the analysis window. Thus, a low value leads to a higher number of analysis frames. (2) A maximum number of formants, which determines the number of expected formants in the calculated spectrum, which are represented in the calculation in form of filter poles. (3) A frequency ceiling (in Hz) for the range of formant estimation. (4) A window length that determines the effective duration (in s) of the analysis window. (5) A formant bandwidth, which determines the frequency range of a single formant frequency. (6) A cut off frequency for pre-emphasis (in Hz; 6 dB amplitude enhancement per octave above this frequency). In the case of 'To LPC...' and its sub-commands, the so-called Nyquist frequency, which is equal to half the sampling frequency of the particular signal, is automatically used as their frequency ceiling for formant estimation. Therefore, this requires (in most cases) resampling

the sound before doing an analysis. This is necessary, because the estimation of, for example, five formants below 5500 Hz requires a sampling frequency of 11 kHz (for more details regarding sampling frequency values and sub-commands of 'To LPC ...' see Praat manual). Among the 'To Formant...' commands, there are several algorithms that can be used for formant estimation: (1) 'To Formant (sl) ...', which is based on the implementation of the Split Levinson algorithm by Willems (1986) that will always find the requested number of formants in every analyzed frame. (2) 'To Formant (keep all)...', which is based on a calculation that keeps all formant values, even those below 50 Hz and those above the frequency ceiling. (3) 'To Formant ...(burg)' allows the investigator to choose all above-mentioned parameter settings manually (the algorithm is described in Childers and Kesler, 1978, and Press et al., 1992). It resamples the sound to a value twice the number of the selected frequency ceiling, applies a pre-emphasis, and computes a number of filter poles that are twice the number of the expected number of formants. (4) To Formant...(robust) is based on the Burg method, but iteratively refines the calculated formant frequencies and bandwidths by selectively weighting sample values (for detailed explanation see Lee, 1988).

5.4.2 Choosing algorithm and parameter settings for linear prediction in Praat

When carrying out standard formant estimation techniques, 'To LPC...' should typically be avoided, because the Nyquist frequency determines the frequency range for the estimation, and, thus, resampling is in most of the cases necessary. But also within the algorithms assembled in 'To Formant ...' commands there are more and

less useful ones for formant estimations. For example, 'To Formant (sl)...' should be handled carefully as it always finds the requested number of formants in every analyzed frame. Furthermore, it has no analysis of the formant bandwidth implemented and applies a default range of 50 Hz. 'To Formant (keep all)...' seems to be critical in terms of F_1 and F_2 identification due to a calculation which keeps all formant values. 'To Formant (burg)...' as well as 'To Formant (robust)...' allows the user to control most of these drawbacks. As for linear prediction required, the user can select numeric values for time step(s), window length(s) of the analysis, a maximum number of expected formants, a frequency ceiling for the analysis, and a cut-off frequency for pre-emphasis. Therefore, both methods seem to be a reasonable technique for LP within Praat. However, the selected parameter settings for both methods need to be chosen carefully, because some of them also rely on speaker specific properties and thus should be adapted for individual speakers. For example, the frequency ceiling (or 'maximum formant') is of crucial importance as it is relatively speaker specific. This is due to the fact that it is based on the size of the vocal tract of the particular speaker (only consider that children have smaller vocal tracts and therefore wider spread resonance frequencies). Therefore, the Praat manual suggests average values for adult males (5000 Hz), adult females (5500 Hz), and children (8000 Hz). The default setting, which is set to 5500 Hz, has, therefore, to be changed to a frequency range corresponding to the age and gender of the specific speaker. As indicated, it is expected for the vocal tract filter to have a speech-related resonance structure only within a specific frequency range (e.g., 5500 Hz for an adult female). Along with analyzing an appropriate frequency range the investigator also should choose

a reasonable number of expected resonances (formants) in that spectrum. Praat suggests searching for five formants within all recommended frequency ranges for adults and children (Note that filter poles of LP analysis are equal to twice the indicated maximum number of formants). However, several researchers (e.g., Ladefoged, 2003, p. 122–125) recommend including an additional formant (i.e. two additional poles) to account for higher formants that may be influencing the spectrum or a pole due to the glottal pulse shape. Another problem concerning frequency ranges evolves with respect to the bandwidth of formants (a frequency region in which the amplification differs less than 3 dB from the amplification at the center frequency; Kerstens et al., 2001). It is a matter of debate, what maximum formant bandwidth might be reasonable to set with regard to the analysis of speech-related formant patterns. Simply put, and not differentiating lower and higher formants, according to Fulop (2011), formant bandwidths > 300 Hz make formant tracking unreasonable.

5.4.3 Spectrographic depiction in Praat

To create a spectrogram in Praat the investigator has to select values for different parameters: (1) Window length(s) that determine(s) the effective duration (in s) of the analysis window[1] . Previous studies have shown that ranges between 15 and 50ms have been profitably used (Fulop, 2011). Furthermore, the window has to

[1] Note that the entered value does not represent the actual duration of the analysis window as Praat doubles this value for a Gaussian-like analysis (for more information see Praat manual).

be at least as long as one glottal pulse period in order to get a reasonable spectrum estimate. When vowels are not steady state, it is also necessary to keep the window short enough to avoid extensive variation within it. (2) A maximum frequency, which determines the frequency range of the spectrogram. This value should be chosen regarding the dependencies to speaker specific properties described above. (3) Time step(s) to determine the frames within the total duration of the analysis window. (4) A frequency step that regulates the resolution. Therefore, lower values result in higher resolutions. Furthermore, in Praat commonly a Gaussian window is used for creating a spectrogram. In contrast to LP analysis, the spectrographic depiction is often considered as straightforward because it possesses high face validity (Ciocca and Whitehill, 2013; for a comparison in terms of accuracy, see Monsen and Engebretson, 1983). For example, a wide-band spectrogram enables, in most cases, visual estimation of formant frequencies (similar to peak-picking from short-term spectra).

5.4.4 Crosschecking within a lot of samples

Regarding the setting of parameters related to these criteria, an investigator must refer to phonetic knowledge in order to rely the parameter settings to the age and gender of the speaker as well as to the expected number of formants and corresponding frequency ranges of the vowel qualities in question, in full understanding of the acoustic meaning and implications of the different parameters for the analysis. Subsequent to the numerical analysis, an investigator should cross check the results on the basis of a spectrogram. Therefore, Praat allows the user to draw the

calculated values from the LP analysis into a spectrogram. If there is no clear relationship between the numerical calculation and the spectrographic depiction, it is recommended to run a new analysis with different parameter settings, above all with a higher or lower maximum of formants for LP analysis. If within such correction, calculated numerical values and spectrographic inspection still differ, in general, the corresponding sounds are excluded from statistical analysis. Both LP and spectrographic depiction can be combined to increase the validity of formant estimation. However, in terms of a high number of samples, this might be extremely laborious. Furthermore, previous studies have described two phenomena, i.e. formant merging and the appearance of 'spurious' formants, which makes cross-checking in terms of formant estimation techniques even more complicated as both methods fail in calculating reasonable formant frequency estimations.

5.4.5 Formant merging and 'spurious' formants

A merging of two formants has been described in a number of previous studies (see, for example, Ladefoged, 2003) and appears often when analyzing a low back vowel. In this case, the spectra show only one prominent peak below 1 kHz. However, according to vowel statistics, there should be two formants. Moreover, a cross check with the spectrogram may be problematic because only one dark bar may be seen, and not two separated bars (for an illustration see Figure 5–1). This phenomenon seems to be primarily caused by changes in the configuration of the vocal tract (e.g., shortening the pharynx as it is assumed to be of importance for F_1; Sundberg et al., 2012). The second serious problem that can cause significant difficulties in terms

of formant estimation is the appearance of so-called 'spurious' formants. Previous studies (see, for example, Ladefoged, 2003) report that an extra formant can appear near the first formant, or make it look like a formant is split into two. In this case, the dark bar in the spectrogram is wider, and it is impossible to say whether the additional energy is above or below the 'genuine' first formant. Ladefoged (2003, p. 114–115) also found, that additional energy can sometimes be observed around the 1 kHz region, irrespective of the vowel. In addition, further studies indicate that a spurious formant can occur in the region between the second and the third formant (e.g., for /ɑ/), or in the region of 1.7–1.8 kHz (Peterson, 1961). In such cases, it is often difficult or even impossible to distinguish between genuine and spurious formants.

5.5 Problems of formant estimation at higher f_o

Another problem that may occur in the acoustic analysis of vowels is the extensive variation of f_o. Existing formant statistics document formant patterns only for f_os typical for 'normal speech' and do not include substantial variation. Formant patterns in connection with different fundamental frequencies mostly only relate to the comparison of different speaker groups, i.e. children, women, and men (e.g., Peterson and Barney, 1952; Hillenbrand et al., 1995). Numerous studies have, in fact, investigated acoustic characteristics of vowel sounds that include an extensive variation of f_o in the context of singing and acting (e.g., Sundberg and Skoog, 1997; Joliveau et al., 2004; Garnier et al., 2010). Further results of acoustic analysis of sounds that include a high variation of f_o are reported from studies on strong emotional expressions, like shouting (e.g., Traunmüller, 1988) and crying (Murry et al.,

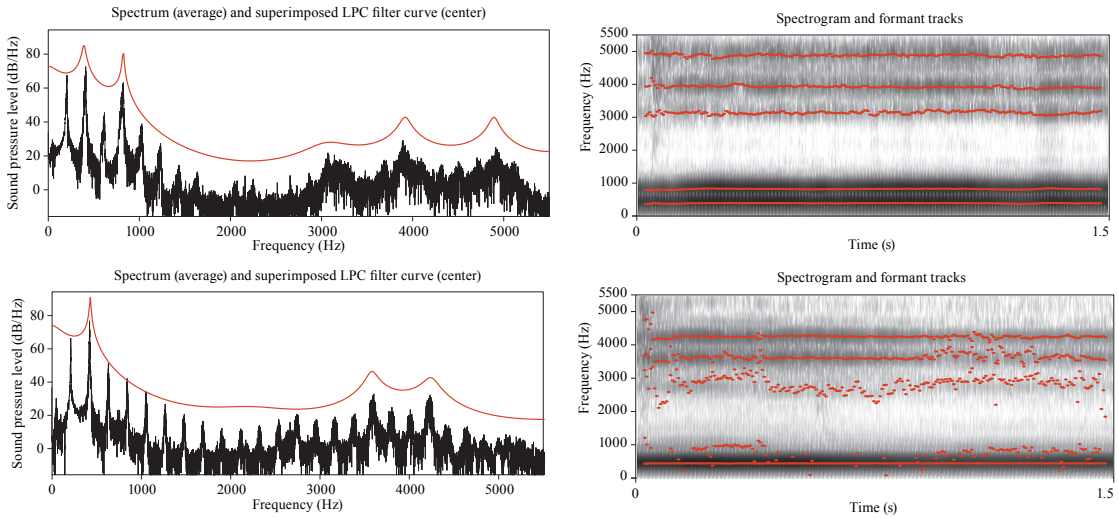

Figure 5–1: Spectra and spectrograms of two sounds of the German closed vowel /o/ at a fundamental frequency of 203 Hz (top rows) and 210 Hz (bottom rows) respectively, produced as maintained isolated sounds by two unprofessional female adult speakers. Praat formant analysis using the Burg algorithm and applying standard parameters for women (analyzed frequency range = 5500 Hz, maximum of formants = 5, window length = 0.025 s, pre-emphasis = 50 Hz) provides two formant frequencies < 1 kHz for the first utterance according to what is expected from phonetic theory and from given values in formant statistics for German vowels (Pätzold and Simpson, 1997), and the inspection of the spectrogram confirms such a result. However, for the second utterance, the same method of formant analysis provides only one lower formant, and the inspection of the spectrogram does also not allow for a determination of a second formant < 1 kHz.

1977), and from studies on vowel synthesis (e.g., Sundberg, 2006).

Our observations have shown that an extensive variation of f_o, exceeding one octave, also appears in everyday life as a characteristic of normal speech (Figure 5–2), so does not only concern artistic, interpretative, and entertaining utterances.

If it holds true that vowels are fairly intelligible at higher f_o, the question arises, how listeners can categorize acoustic realizations of vowels at varying fundamental frequencies. This question becomes even more interesting when considering that, for example, female speakers easily produce vowels up to 500 Hz, and thus, utterances, where f_o exceeds values found for several F_1 in previous studies (e.g., /i, y, u/, for which F_1 is around 350 Hz for German vowels produced by female speakers; see Pätzold and Simpson, 1997; see Fig. 5–3, for an example). Regarding the difficulties of formant estimations at higher fundamental frequencies it is not surprising that previous studies indicated that reasonable formant estimation is limited up to a certain frequency. For example, Monsen and Engebretson (1983) argued that the accuracy of both linear prediction and spectrographic analysis decreases greatly when the fundamental frequency is 350 Hz or above (which, as mentioned, often has been found as an average frequency of F_1 for /i, y, u/; Pätzold and Simpson, 1997). Other studies even argued that formant frequency estimation is only reasonable for $f_o < ½\ F_1$ (e.g., Thalén and Sundberg, 2001).

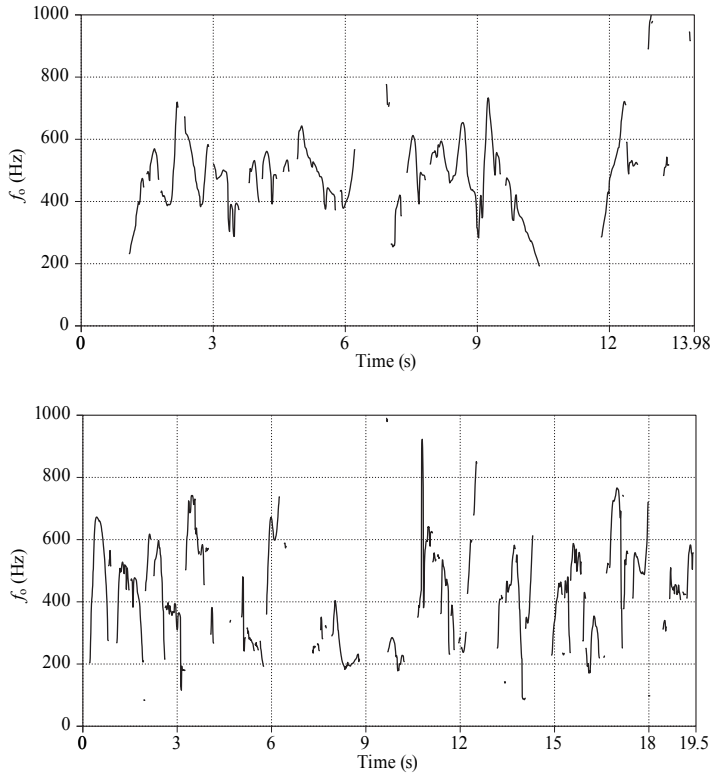

Figure 5–2: Fundamental frequency contours: (top) French native speaker (adult female) selling grilled chicken on a market in Paris; (bottom) French native speaker (adult female, Comedienne) during an appearance on a French television show. Our observations have shown that f_o can easily reach values around 500 Hz, and even above. Both utterances still seem fairly intelligible at such frequencies.

This appears understandable considering, for example, that two or more harmonics define a formant. However, if $f_o > ½\ F_1$, only one harmonic may define a formant, or three harmonics may define two formants. In this case, the frequency distance between the harmonics is already too high. Thus, the resolution of the harmonics is too low to define properly a spectral envelope. Obviously, the definition of a spectral peak is then also problematic.

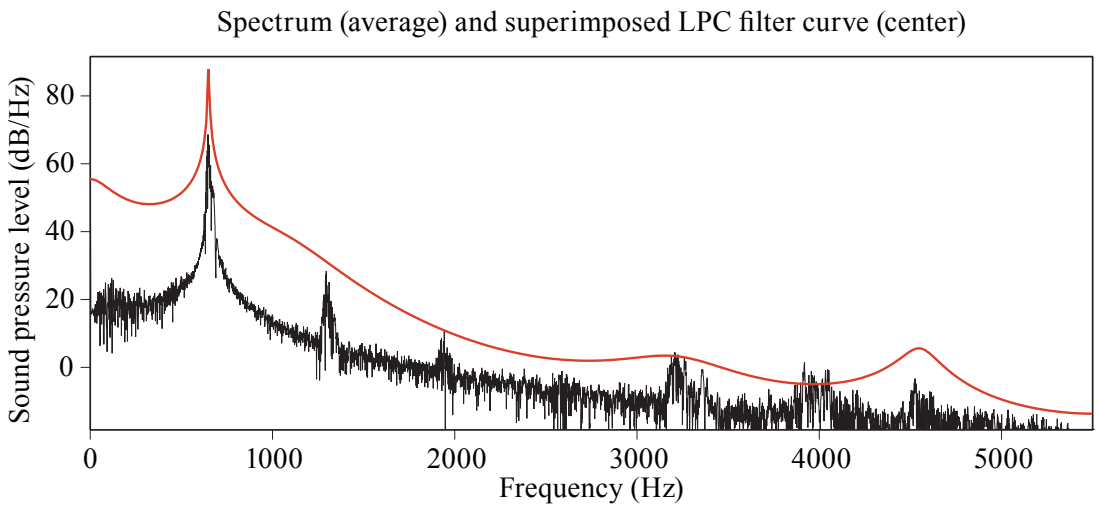

Figure 5–3: Average FFT spectrum of the German vowel /u/ produced by an adult female speaker as maintained isolated sound with f_o exceeding F_1 value given in formant statistics (f_o = 647 Hz; F_1 for the German vowel /u/ of female speakers should be expected around 350 Hz, see Pätzold and Simpson, 1997). An LPC filter curve calculated with the default parameter settings in Praat is drawn above the spectrum.

In terms of identification of single vowel qualities at different f_o, two approaches might be helpful for future studies: (1) Acoustic analysis also apart from formant estimation and including first-hand measurements of the resonance frequencies (see, for example, Joliveau et al., 2004; Wolfe et al., 2009), and (2) vowel (re)synthesis, for which the formant frequencies are calculated in a first step and then used for synthesizing vowel sounds that need to be identified and compared to the 'natural' sounds. Including an extensive variation of f_o in vowel synthesis could also generate new knowledge, as comparing statistical values of 'natural' sounds with sounds produced by one- and two-formant synthesis, differences can already be found for

F_1 and F_1'[2] for back vowels, and for F_2 and F_2' of front vowels (Delattre, 1948; Delattre et al. 1952; Fant and Risberg, 1963; Carlson et al., 1974; see also Bladon and Fant, 1978). Moreover, a direct correspondence of formant patterns and vowel qualities is only demonstrated for sounds not exhibiting extensive variation of f_o.

5.6 Acknowledgments

This work was supported by the Swiss National Science Foundation SNSF Grant No. 100016_143943/1.

5.7 References

Baken, R. J., & Orlikoff, R. F. (**2000**). Clinical measurement of speech and voice. Clifford Park NY: Delmar, Cengage Learning.

Bele, I.V. (2006). The speaker's formant. Journal of Voice **20**(4), 555–578.

Bladon, A., & Fant, G. (**1978**). A two-formant model and the cardinal vowels. STL-QPSR, 19(1), 1–8.

Boersma, P. (**1993**). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings **17**, 97–110.

Boersma, P., and Weenink, D. (**2014**). Praat: Doing phonetics by computer. Version 5.3.62. (retrieved from www.praat.org).

[2] The apostrophe serves here as an indicator for the statistical value in order to distinguish it from the actual values found for a speaker.

Carlson, R., Fant, G., and Granström, G. (**1974**). Two-formant models, pitch and vowel perception. Acustica, 31(6), 55–82.

Chiba, T., and Kajiyama, M. (**1941**). The vowel: It's nature and structure. Tokyo: Kaiseikan.

Childers, D. G., and Kesler, S. B. (**1978**) (Eds.), Modern spectrum analysis. Vol. 331. New York: IEEE Press.

Ciocca, V., and Whitehill, T. L. (**2013**). The Acoustic Measurement of Vowels. In: M. J. Ball, and F. E. Gibbon (Eds.), Handbook of Vowels and Vowel Disorders (pp. 113–137). New York, London: Psychology Press, Taylor and Francis Group.

Cleveland, T. F., Sundberg, J., and Stone, R. E. (**2001**). Long-term-average spectrum characteristics of country singers during speaking and singing. Journal of Voice **15**(1), 54–60.

Delattre, P. (**1948**). Un triangle acoustique des voyelles orales du français. The French Review **21**(6), 477–484.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (**1952**). An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. WORD **8**(3), 195–210.

Fant, G. (**1959**). Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics No. 1: 1–106.

Fant, G. (**1960**). Acoustic theory of speech production. The Hague: Mouton & Co.

Fant, G., and Risberg, A. (**1963**). Auditory matching of vowels with two formant synthetic sound. STL-QPSR **4**(4), 7–11.

Fulop, S. A. (**2011**). Speech spectrum analysis. Berlin/Heidelberg: Springer.

Garnier, M., Henrich, N., Smith, J., and Wolfe, J. (**2010**). Vocal tract adjustments in the high soprano range. Journal of the Acoustical Society of America **127**(6), 3771–3780.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). Acoustic correlates of breathy vocal quality. Journal of Speech and Hearing Research **37**, 769–778.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). Acoustic characteristics of American English vowels. Journal of the Acoustical Society of America **97**(5), 3099–3111.

Imagawa, H., Sakakibara, K., Tayama, N., and Niimi, Seiji (**2003**). The effect of the hypopharyngeal and supra-glottic shapes on the singing voice. Proceedings of the Stockholm Music Acoustics Conference, 2003, Stockholm, Sweden, 3–6.

Joliveau, E., Smith, J., and Wolfe, J. (**2004**). Vocal tract resonances in singing: The soprano voice. Journal of the Acoustical Society of America **116**(4), 2434–2439.

Kent, R. D., and Read, C. (**2002**). The Acoustic Analysis of Speech. Clifton Park, NY: Delmar, Cengage Learning.

Kerstens, J. Ruys, E., and Zwarts, J. (**2001**). Lexicon of Linguistics. http://www2.let.uu.nl/uil-ots/lexicon/ (accessed 01/03/2014).

Ladefoged, P. (**2003**). Phonetic data analysis: An introduction to fieldwork and instrumental techniques. Malden, MA: Blackwell.

Lee, C.-H. (**1988**). On robust linear prediction of speech. IEEE Transactions on Acoustics, Speech, and Signal Processing **36**(5), 642–650.

Mathews, M. (**1999**). Introduction to timbre. In: P. R. Cook (Ed.), Music, cognition, and computerized sound (pp. 79–87). Cambridge: MIT Press.

Monsen, R. B., and Engebretson, A. M. (**1983**). The accuracy of formant frequency measurements: A comparison of spectrographic analysis. Journal of Speech and Hearing Research, 26, 89–97.

Murry, T., Amundson, P., and Hollien, H. (**1977**). Acoustical characteristics of infant cries: fundamental frequency. Journal of child language **4**, 321–328.

Pätzold, M., and Simpson, A. (**1997**). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. Arbeitsberichte des Instituts für Phonetik und digitale Spachverarbeitung Universität Kiel **32**, 215–247.

Peterson, G. E., and Barney, H. L. (**1952**). Control methods used in a study of the vowels. Journal of the Acoustical Society of America **24**(2), 175–184.

Peterson, G. E. (1961). Parameters of vowel quality. Journal of Speech and Hearing Research, 4, 10–29.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (**1992**). Numerical recipes in C: The art of scientific computing. Cambridge: Cambridge University Press.

Stevens, K., and House, A. (**1955**). Development of a Quantitative Description of Vowel Articulation. Journal of the Acoustical Society of America **27**(3), 734–742.

Sundberg, J., and Skoog, J. (**1997**). Dependence of jaw opening on pitch and vowel in singers. Journal of Voice **11**(3), 301–306.

Sundberg, J. (**2006**). The KTH synthesis of singing. Advances in Cognitive Psychology **2**(2–3), 131–143.

Sundberg, J. (**2012**). Perception of singing. In: Deutsch, D. (Ed.), The psychology of music (pp. 69–106). Waltham, MA: Academic Press.

Sundberg, J., Lã, F. M. B., and Gill, B. P. (**2011**). Professional male singers' formant tuning strategies for the vowel /a/. Logopedics, phoniatrics, vocology **36**(4), 156–167.

Thalén, M., and Sundberg, J. (**2001**). Describing different styles of singing. A comparison of a female singer's voice source in 'Classical', 'Pop', 'Jazz' and 'Blues'. Logopedics, phoniatrics, vocology **26**(2), 82–93.

Traunmüller, H. (**1988**). Paralinguistic variation and invariance in the characteristic fre- quencies of vowels. Phonetica **45**(1), 1–29.

Willems, L. F. (**1986**). Robust formant analysis. IPO Annual report **529**, 1–25. Eindhoven: Institute for Perception Research.

Wolfe, J., Garnier, M., and Smith, J. (**2009**). Vocal tract resonances in speech, singing and playing musical instruments. Human Frontier Science Program Journal **3**, 6–23.

Appendix

Symbolic notation

Recently, an attempt was made by a group of voice scientists (Titze et al., 2015) to reach some consensus on the symbolic notation of the terms *fundamental frequency*, *formant frequency*, *resonant frequency*, and *harmonic*. In this thesis, I follow their suggestion and use f_o (lower case "f" followed by a subscript "o" as an abbreviation of "oscillation") as the symbol for fundamental frequency, and I consider all harmonics as multiples of f_o, i.e., f_o, 2 f_o, 3 f_o, ..., n f_o. Formant frequencies are abbreviated as F_1, F_2, F_3, ... , F_n (capital "F" followed by a subscript number indicating the order). As symbolic notation for resonant frequencies, I use f_{R1}, f_{R2}, f_{R3}, ... , f_{Rn} (lower case "f" followed by a subscript "R" and a number indicating the order).

Fundamental frequency and pitch

In speech communication research, the terms *pitch* and *fundamental frequency* are often used interchangeably. However, the pitch is originally understood as the term representing the perceptual sensation that derives from an acoustical feature of the speech signal, i.e., the fundamental frequency. Pitch is usually perceived as low when the fundamental in the spectrum is low as well. However, previous studies have also shown that pitch perception does not necessarily change when, for example, the first

harmonic (i.e., the fundamental) is canceled out from the signal, but the spacing of the harmonics stays the same (Smith et al., 1978). Therefore, both terms are closely connected, but should not be used synonymously except the meaning and interpretation of the term is obvious.

Formants and resonances

In some studies, the terms *formant* and *resonance* are used synonymously. As this might be confusing, I follow the definitions provided by ANSI and ASA and distinguish between the two. In the current *American National Standard Acoustical Terminology* published by the American National Standards Institute, Inc. and the Acoustical Society of America a *formant* is described as a "range of frequencies in which there is an absolute or relative maximum in the sound spectrum" (ANSI/ASA S1.1–2013:p.62). The frequency at the peak or maximum amplitude of a formant is referred to as *formant frequency*. A *resonance* is defined as a "phenomenon that exists for a linear system in harmonic forced oscillation when any change in the excitation frequency results in a decrease in the response of the system", and a *resonance frequency* is regarded as the "frequency at which resonance exists" (ANSI/ASA S1.1–2013:p.18).

Harmonics and overtones

In this thesis, a harmonic is considered as a sinusoidal quantity with a frequency, which is an integral multiple of the fundamental frequency (ANSI/ASA S1.1–2013, p. 4). In previous research, the term "overtone" has often been used instead, although it does not include the fundamental (i.e., the second harmonic being called the first

overtone). Following the recommendation of the *Accredited Standards Committee S1, Acoustics* in the *American National Standard Acoustical Terminology* (ANSI/ASA S1.1–2013, p. 63), I do not use the term overtone in this dissertation.

References

ANSI (**2013**). ANSI/ASA S1.1–2013, Acoustical Terminology (American National Standards Institute, Inc., New York).

Smith, J. C., Marsh, J. T., Greenberg, S., and Brown, W. S. (**1978**). "Human auditory frequency-following responses to a missing fundamental," Science **201**(4356), 639–641.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., Švec, J. G., Ternström, S., and Wolfe, J. (**2015**). "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," J. Acoust. Soc. Am. **137**(5), 3005–3007.

Curriculum Vitae

Name: Friedrichs
First name: Daniel
Birthdate: 29 November 1980
Birthplace: Osnabrück (Germany)
✉ daniel.friedrichs@ucl.ac.uk
☎ +41-78-7092911
ORCID ID: 0000-0003-1102-4916

Higher Education

3/2013–8/2016	**Ph.D. Studies in General Linguistics (summa cum laude)** University of Zurich (Switzerland) Ph.D. program Linguistics ("Linguistic Structure Linguistic Variation Linguistic History") Thesis: Beyond Formants: Vowel Perception at High Fundamental Frequencies Advisors: Prof. Dr. Volker Dellwo and Prof. Dr. Martin Meyer **Date of Ph.D. viva: 29.8.2016**
1/2015	**Lehrdiplom für Maturitätsschulen (Subjects: German and History)** University of Zurich (Switzerland)
9/2012	**M.A. in German Language and Literature and General History** University of Zurich (Switzerland) Thesis: Variabilität akustischer Rhythmuskorrelate bei der Sprechersynchronisierung (Grade: 6/6)
11/2010	**B.A. in German and History (with teaching option)** Humboldt University of Berlin (Germany)

Current and previous positions

since 9/2016	**Post-doctoral Research Fellow (Early Postdoc.Mobility Fellowship)** Department of Speech, Hearing and Phonetic Sciences University College London (United Kingdom)
3/2015–8/2016	**Ph.D. Candidate and Research Associate (Forschungskredit Candoc UZH)** Phonetics Laboratory, Department of Comparative Linguistics University of Zurich (Switzerland)
9–12/2015	**Visiting Ph.D. Student** Department of Speech, Hearing and Phonetic Sciences University College London (United Kingdom)
3/2013–2/2015	**Ph.D. Candidate and Research Associate (SNSF)** Phonetics Laboratory, Department of General Linguistics University of Zurich (Switzerland)
6/2012–2/2013	**Research Assistant** Phonetics Laboratory, Department of General Linguistics University of Zurich (Switzerland)

London, 2017